A Quality-Assurance Study of ChEBI

Hasan Yumak, Ling Chen
BMCC, CUNY
New York, NY USA
{hyumak, lchen}@bmcc.cuny.edu

Michael Halper, Ling Zheng, Yehoshua Perl, Gai Elhanan
New Jersey Institute of Technology
Newark, NJ USA
{michael.halper, lz265, perl, gai.elhanan}@njit.edu

***Abstract*—Ontologies are important components of many health-information systems. The Chemical Entities of Biological Interest (ChEBI) ontology has become a standard reference for chemicals appearing in biological contexts. As such, assuring the quality of its content is imperative. In fact, ChEBI has a dedicated Web page at which errors and inconsistencies in its concepts can be reported. A study of the correctness of a random sample of ChEBI concepts is carried out. The results show that quite a large number of ChEBI concepts suffer from some kind of problematic modeling. For example, we found that 15.5% of the sample concepts exhibited severe errors of commission, including incorrect hierarchical (*is a*) and lateral relationships. Errors of omission were also prevalent. The overall results of our quality-assurance (QA) study are presented. Suggestions for enhancing the QA processes in place for ChEBI are discussed.**

Keywords—ChEBI; chemical ontology; chemical concept; quality assurance; modeling error; error distribution

I. INTRODUCTION

Ontologies are structures that capture terminological knowledge for some target domain. Typically large in size and high in complexity, ontologies have become fundamental fixtures of health and biological information processing environments. The Chemical Entities of Biological Interest (ChEBI) ontology [1] is an authoritative reference that models chemical concepts having biological significance, particularly from the perspectives of molecular structure and biological role or application [2]. Maintained by the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), it is an important chemical annotation and identification standard. As of its February 2016 release, it comprised a collection of 61,895 concepts (including 47,752 fully annotated compounds), 104,351 *is a* (hierarchical) relationships and 65,077 lateral relationships.

Due to their scope and complexity, it is nearly impossible for ontologies such as ChEBI to be free of modeling errors and inconsistencies. This can hinder their usefulness and adversely affect the software systems and applications dependent on them. ChEBI has been used, for example, as a source for annotations in various bioinformatics databases, including UniProt which utilized ChEBI for the cofactor comments related to enzymes [3]. Metabolites in a human metabolism model have been annotated with terms from ChEBI [4]. ChEBI is also used to support text mining and chemical analysis. For example, in a recent study [5], a novel method for computing semantic similarity between chemical entries based on ChEBI was introduced to improve the chemical entity identification in texts. In another study [6], a new prediction method based on information from ChEBI for identifying drugs' target groups was proposed. ChEBI's structural hierarchy has been integrated with the Gene Ontology (GO) [7] to allow for data integration across the biology and chemistry domains. Errors in ChEBI could be propagated in a deleterious manner into such applications. Due to this, assuring the quality of the conceptual content of an ontology is a very important matter. ChEBI, in fact, employs a GitHub issue tracking system (https://github.com/ebi-chebi/ChEBI/issues) to enable users to report various errors and inconsistencies that they encounter while using the ontology. Those reports are handled by ChEBI's curators.

In this paper, we are interested in assessing the percentage of ChEBI's concepts that suffer from some type of modeling issues. We refer to concepts that have errors or inconsistencies in their modeling as being *erroneous* concepts. We selected a random sample of 400 concepts for our study. Two subject-domain experts in the field of chemistry were asked to carry out separate quality-assurance (QA) analyses of the entire sample and then produce a consensus report. The findings revealed that a substantial number of ChEBI concepts suffer from errors of commission and omission, suggesting the need for a formal initiative to map out an ongoing QA project for improvement of the quality of the modeling in ChEBI. Further recommendations for such a project are discussed.

II. METHODS

A QA study of a random sample of ChEBI concepts was carried out. ChEBI is updated monthly, and the version we used in this study was that of February 2016, which contained 61,895 concepts, including 47,752 fully annotated compounds. The sample was chosen using the basic sampling technique, simple random sampling without replacement [8]. The sampling frame was all 61,895 concepts in the February 2016 release. A random number drawn from a uniform distribution over the range [0, 1] was generated as a key for each concept. All concepts were sorted using the keys, and the smallest 400 concepts were selected as the random sample [9].

It should be noted that ChEBI employs three hierarchies to classify molecular entities. Its chemical entity hierarchy, the largest with 60,537 concepts (97.8% of all ChEBI concepts), is used to classify molecular entities according to their chemical structure. The subatomic particle hierarchy with 42 concepts categorizes particles smaller than atoms. ChEBI's third hierarchy, the role hierarchy with 1,322 concepts, itself has three subhierarchies that define the roles in different contexts

for the compounds, namely, the application subhierarchy to represent the intended use by humans for the compounds (e.g., fuel and anti-inflammatory agent), the biological role subhierarchy to represent the roles of compounds within the biological context (e.g., growth regulator and inhibitor), and the chemical role subhierarchy (e.g., acid and base). Note that there are six concepts belonging to both the chemical entity and subatomic particle hierarchies, e.g., *helion* and *proton*. The concepts randomly selected for our study came from all three of the main hierarchies, irrespective of hierarchy.

The QA analysis was done by a pair of chemistry subject-domain experts. In the initial step, each concept from the sample was inspected by each of the experts separately—without any communication between them. Their results were tabulated in two individual error reports. Within a report, the rationale for the judgment of any error was recorded, and a suggested correction was proffered. Afterward, a combined report was prepared listing the respective findings of both experts for all the concepts. This report was shared with both experts who were then each asked separately to mark their agreement or disagreement with the findings of the other person—and to review their own findings in light of the other's. After a review of the other's report, each expert was able to change their mind regarding their own judgment of a modeling error. A concept previously judged to be modeled in error could instead be deemed to be correct, and vice versa. After this step of consensus building, a concept was deemed to be *erroneous* if both subject-domain experts agree that it was such.

Most of the QA analysis of ChEBI centered on issues with concepts' relationships. The three primary relationships in ChEBI are the hierarchical *is a* relationship, capturing standard subsumption in hierarchies, the relationships *has part*, indicating the whole/part association between compounds, and *has role*, linking concepts in the chemical entity hierarchy to concepts in the role hierarchy. There are seven chemistry-specific lateral relationships, namely, *is conjugate base of*, *is conjugate acid of*, *is tautomer of*, *is enantiomer of*, *has functional parent*, *has parent hydride*, and *is substituent group from*. In combination, the relationships in ChEBI can form converses. For example, if concept *A is conjugate base of* concept *B*, then *B is conjugate acid of* A. A similar situation exists for the two relationships *is tautomer of* and *is enantiomer of* [10].

The types of errors that the experts were looking for included both *errors of commission* and *errors of omission*. Examples of the former are incorrect hierarchical relationship, incorrect lateral relationship, and incorrect relationship target. Examples of the latter are missing hierarchical relationship and missing lateral relationship.

III. Results

A random sample of 400 concepts (0.6%) was selected from the 61,895 concepts in ChEBI, February 2016 release. Out of these 400 concepts, 388 (97%) are from the chemical entity hierarchy, 11 are from the role hierarchy, and one is from the subatomic particle hierarchy. In the following, we often refer to a ChEBI concept using its name along with its unique ChEBI id, written, for example, as "CHEBI: 31900" (which is the concept with the name *Neticonazole hydrochloride*).

Two of the authors (HY and LC), both subject-domain experts in chemistry and experienced in ontology QA, carried out the individual QA analyses on the entire sample of concepts and then produced a consensus report on the errors discovered. Out of the 400 concepts, the two subject-domain experts agreed on the errors reported for 167 concepts (41.8%). The margin of error at the 95% confidence level for a 400 concept sample from a population of 61,895 concepts is 4.9% [11]. Among the 167 erroneous concepts, 166 of them are from the chemical entity hierarchy, and the other is from the role hierarchy.

There were 122 (30.5% = 122 / 400) concepts that exhibited errors of omission. Of these, 121 concepts were found to be missing hierarchical relationships, and only one concept, *fatty acid anion 4:0* (CHEBI: 78115), was reported to be missing the relationship *is conjugate base of*.

Table 1 shows the number and percentage of concepts with errors of commission. For example, 36 concepts (9%) in the sample were found to have incorrect *is a* relationships. Note that some concepts may have multiple kinds of errors. For example, there are 17 concepts that were reported to have both errors of commission and omission.

Table 2 and Table 3 list examples of erroneous concepts with errors of commission and omission, respectively, along with their corresponding suggested corrections and the reason for the error. For example, in Table 2 (Row 3), we see that *Neticonazole hydrochloride* (CHEBI: 31900) was originally modeled as *is a hydrochloride*; instead, the modeling should be *has part* because the mixture contains hydrochloride.

TABLE 1. Distribution of erroneous concepts with errors of commission

Error Type	# Erroneous Concepts	% (/400)
Incorrect hierarchical relationship	36	9%
Incorrect lateral relationship	1	0.25%
Incorrect relationship target	28	7%
Total:	**62**	**15.5%**

TABLE 2. Examples of concepts with errors of commission

Error & Correction Type	Concept (ChEBI ID)	Correction	Reason
Incorrect hierarchical relationship removed	*3beta,13-Dihydroxy-16-(hydroxymethylene)-13,17-seco-5alpha-androstan-17-oic acid, delta-lactone* (CHEBI:79677)	Remove CHEBI:26979 *organic heterotricyclic compound* Remove CHEBI:51959 *organic tricyclic compound* Remove CHEBI:36688 *heterotricyclic compound*	It has more than 3 cyclic structures.
Incorrect hierarchical relationship replaced	*all-trans-polyprenyl diphosphate* (CHEBI:55337)	Replace CHEBI:37531 *polyprenol diphosphate* with CHEBI: 26248 *prenyl group*	Prenol is an alcohol and prenyl is an alkene.
Incorrect relationship replaced	*Neticonazole hydrochloride* (CHEBI:31900)	Change the relationship *is a* CHEBI:36807 *hydrochloride* to *has part hydrochloride*	Hydrochloride (HCl) is a part of the mixture, but the concept itself is not HCl.
Incorrect relationship target replaced	*(S)-3-hydroxyoctanoyl-CoA(4−)* (CHEBI:62617)	Change the charge of its conjugate from 0 to 3−	Conjugate acid and its base should be different by only 1 charge due to loss or gain of one proton.
Incorrect relationship target replaced	*alkane-alpha,omega-diammonium(2+)* (CHEBI:70977)	Change the charge of its conjugate from 0 to 1+	Conjugate acid and its base should be different by only 1 charge due to loss or gain of one proton.

TABLE 3. Examples of concepts with errors of omission

Error & Correction Type	Concept (CHEBI ID)	Correction	Reason
Missing hierarchical relationship added	*2-hydroxydibenzofuran* (CHEBI:34287)	Add CHEBI: 33836 *benzenoid aromatic compound*	It has two bezene rings.
Missing hierarchical relationship added	*1,2,3,7,8-Pentachlorodibenzofuran* (CHEBI:81507)	Add CHEBI: 28097 *chlorobenzene* Add CHEBI: 33836 *benzenoid aromatic compound*	There are five chlorine atoms bond to two benzene rings.
Missing relationship added	*fatty acid anion 4:0* (CHEBI:78115)	Add the relationship *is conjugate base of* CHEBI:35366 *fatty acid*	Its conjugate is fatty acid, which is not shown.

TABLE 4. Typical errors from the chemistry point of view

Error Type	# Erroneous Concepts	% (/400)
Missing chemical classification	121	30.25%
Incorrect charge difference between conjugate acids and bases	28	7%
Incorrect chemical classification	15	3.75%
Incorrect number of cyclic units	13	3.25%
Incorrect amide classification	7	1.75%
Unmatched chemical name and structure	1	0.25%

Table 4 presents the typical errors found from the chemistry viewpoint along with the numbers of concepts (and sample percentages) exhibiting each kind of error. In the following subsections, we provide detailed analyses of the errors in Table 4. Note that the error classifications in the table are not necessarily disjoint, meaning some concepts may have several kinds of errors. Hence, no totals are provided at the bottom of Table 4.

A. Missing chemical classification

This is the most common error, with 30.25% of the concepts exhibiting it. For example, there are 23 benzene-containing compounds in the sample that should be classified as a *benzenoid aromatic compound* (CHEBI: 33836). Included among these is *8-hydroxy-3-chloro-dibenzofuran* (CHEBI: 79743).

B. Incorrect charge difference between conjugate acid and conjugate base

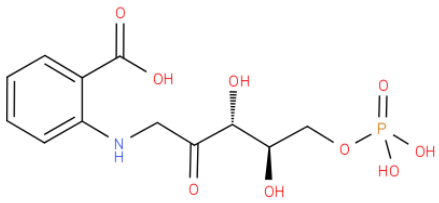
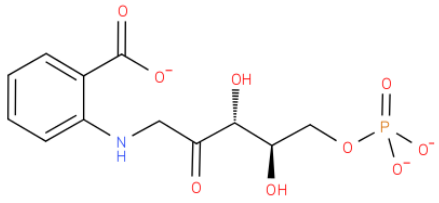
This is the second most common error (7%). The correct charge difference between an acid and its conjugate base is 1, and the acid is 1 charge higher. This is because the acid has one extra proton (H+) compared to its conjugate base. As indicated in the equation HA → A− + H+, HA is the conjugate acid of A−, and A− is the conjugate base of HA. The only difference between HA and A− is one proton; thus, HA is 1 charge higher than A−. For example, ChEBI concept *1-(2-carboxyphenylamino)-1-deoxy-D-ribulose 5-phosphate* (CHEBI: 29112) has as its conjugate base *1-(2-carboxylatophenylamino)-1-deoxy-D-ribulose 5-phosphate (3−)* (CHEBI: 58613). However, its conjugate base charge should be 1−, not 3−, since only one proton is removed from the acid, not three protons, as shown in its structure in Table 5.

C. Incorrect chemical classification

Piperidine (CHEBI: 18049) is classified under *Brønsted acid* (CHEBI: 39141), but in fact it is should be classified as a *Brønsted base* (CHEBI: 39142). This is because an amine is a proton acceptor, not a donor, and it acts as a base not as an acid. Another more common occurrence of this kind of error is seen for 14 erroneous concepts that are classified as some class "*A*," but, in fact, do not have chemical structure *A*. For example, in Fig. 1 showing the chemical structure of *1(3)-O-(alk-1-enyl)-glycerol* (CHEBI: 77998), it does not

contain the following chemical groups *carboxylic ester* (CHEBI: 33308), *carbonyl compound* (CHEBI: 36586), and *ester* (CHEBI: 35701). However, it is classified as a *carboxylic ester*, *carbonyl compound*, and *ester* in ChEBI.

TABLE 5. Structure comparison between conjugate acid and conjugate base

ChEBI Concept	Structure
Acid (CHEBI:29112)	
Conjugate base (CHEBI:58613)	

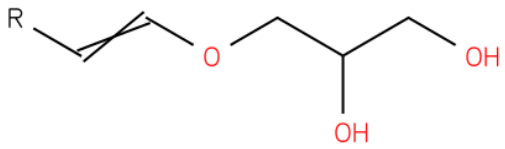


Fig. 1. Structure of *1(3)-O-(alk-1-enyl)-glycerol* (CHEBI: 77998)

D. *Incorrect number of cyclic units*

There are 13 concepts reported with incorrect numbers of cyclic units. For example, *buspirone hydrochloride* (CHEBI: 3224) is classified as an *organic heteromonocyclic compound* (CHEBI: 25693) that has one cyclic structure. However, from the structure seen in Fig. 2, we can see that this concept contains four cyclic units. Similar errors are seen in *(+)-tephrosone* (CHEBI: 66201), *3beta,13-Dihydroxy-16-(hydroxymethylene)-13,17-seco-5alpha-androstan-17-oic acid, delta-lactone* (CHEBI: 79677), *c[G(2',5')pA(3',5')p]* (CHEBI: 75947), *diazoline* (CHEBI: 53123), *dipyridodiazepine* (CHEBI: 63667), *pyrazolopyridazine* (CHEBI: 48383), etc.

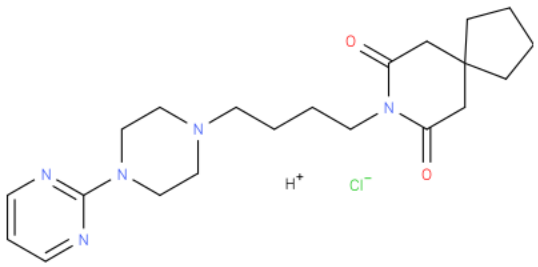


Fig. 2. Structure of *buspirone hydrochloride* (CHEBI: 3224)

E. *Incorrect amide classification*

There are seven concepts classified as *primary amide* (CHEBI: 33256). However, they should be classified as *secondary amide* (CHEBI: 33257). For example, from Fig. 3, we can clearly see that *Arachidonoyl dopamine* (CHEBI: 31231) is a secondary amide (nitrogen group connected two carbon atoms), while it is denoted as a primary amide. Similar errors are seen in *beta-D-glucosyl-(1<->1')-N-eico-sanoylsphinganine* (CHEBI: 84703), *bistratamide I* (CHEBI: 65508), *N-(2 hydroxyhexacosanoyl)phyto-sphingosine* (CHEBI: 64958), *N-(3-oxohexanoyl)homo-serine lactone* (CHEBI: 29640), *N-(2-hydroxy-docosanoyl)eicosasphinganine* (CHEBI: 66983), and *N(4)-acetylcytidine* (CHEBI: 70989).

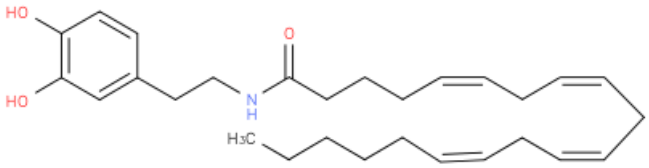


Fig. 3. Structure of *Arachidonoyl dopamine* (CHEBI: 31231)

F. *Name does not match the structure*

As an example, the structure of *diacylglycerol 38:7* (CHEBI: 86986) does not match with its name. Its name indicates that there are two esters, while its structure, seen in Fig. 4, has three R groups, meaning a triacylglycerol, not a diacylglycerol.

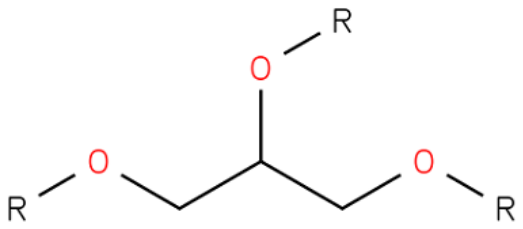


Fig. 4. Structure of *diacylglycerol 38:7* (CHEBI: 86986)

IV. Discussion

In this study, we discovered an error-rate of 42% in a random sample of ChEBI concepts. Among these, 15.5% suffered from severe errors of commission, e.g., incorrect parents and incorrect relationship targets. The remaining 30.5% exhibited errors of omission. (Some concepts had both kinds of errors.) One needs to compare this finding to the reality in other ontologies of a similar caliber. One ontology for which data of this kind exist is SNOMED CT [12]. In several previous studies performed by our SABOC team for evaluating various QA methodologies for SNOMED CT, we were also measuring the percentage of erroneous concepts in random control samples [13-16]. In those studies, we encountered error-rate percentages of 8.3%, 29%, 13%, 8.8%, and 9%, respectively, for the

control samples. Hence, the average control sample error rate was 13.62%. In this light, the results of the present study can be taken to be troubling.

TABLE 6. Distribution of erroneous concepts according to their ChEBI star status

Status	# Concepts Analyzed	# Erroneous Concepts	%
1-star	30	1	3.3%
2-star	78	50	64.1%
3-star	292	116	39.7%

Let us note that ChEBI employs a star status (rating) system to indicate the level of annotation applied to a concept by the ChEBI curatorial team. A concept manually annotated by the team has a "3-star" status. A concept manually annotated by a third party has a "2-star" status. A preliminary concept loaded automatically from a data source but not yet manually annotated is designated with a "1-star" status. We looked at the distribution of erroneous concepts according to their star status (see Table 6). Out of the 400 concepts that we analyzed, 292 concepts had a 3-star status, and 116 of those concepts (39.7%) were deemed to be erroneous. Among the 1-star concepts, 3.3% were erroneous (Table 6). And 64.1% of the 2-star concepts were erroneous. Collectively, the non-3-star concepts exhibited an error-rate of 47.2% (51 out of 108). So, while it was not surprising that the 3-star concepts showed an overall lower error rate, they still contributed significantly to the error findings with a rate of nearly 40%.

In the present study, we assessed the frequency of errors of commission and omission in a random sample of concepts from ChEBI. During our original design and analysis, we postulated that concepts with higher numbers of parents would exhibit higher error rates. This postulation was based on a recurring, substantiated theme in our previous ontological QA research that more complex concepts are prone to exhibit higher error rates than concepts in random control samples. Concepts with multiple parents are indeed more complex than concepts with a single parent due to multiple inheritance of properties and convergence of definitional paths. For example, in our QA research on the CORE problem list of SNOMED CT, we indeed have shown that the expected error rate increases with the number of parents [17].

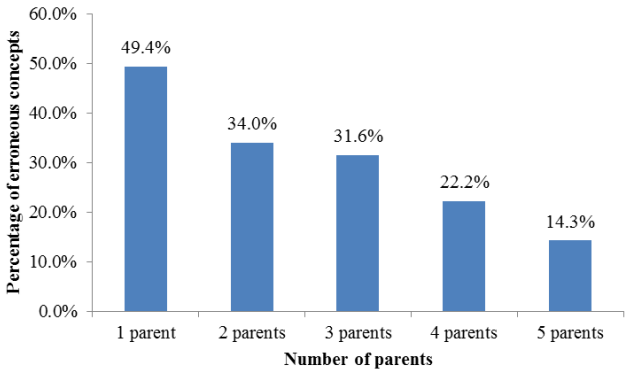


Fig. 5. Distribution of erroneous concepts according to their numbers of parents

However, as seen in Fig. 5—and contrary to our expectations—our postulation was wrong. In fact, our findings show that the error rate is inversely proportional with respect to the number of parents. For example, 12,128 concepts in ChEBI have two parents. From these, 97 were chosen for our random sample; 33 of them (34.0%) were found to be erroneous. Of the 38 three-parent concepts in the sample, 12 (31.6%) were erroneous. Further reductions in the error rates were seen for four-parent (22.2%) and five-parent concepts (14.3%).

ChEBI is user driven, and user requests can be made via the ChEBI submission tool [18]. For a new concept request, users need to provide minimal unique information, including the classifications for the new entity. Users can also report issues or bugs using ChEBI's GitHub issue tracking system (https://github.com/ebi-chebi/ChEBI/issues). As of June 2016, there were 2,933 closed issues and 234 open issues in the tracking system. After ChEBI's curators have verified requests, new concepts and properties are made available in subsequent releases. For example, a user reported on December 11, 2015 that an *is a* relationship should be added between the concepts *endocannabinoid* (CHEBI: 67197) and *lipid* (CHEBI: 18059). On January 28, 2016, a ChEBI curator responded that the change was done. From an inspection of the ontology in its January 2016 version, we can see that *endocannabinoid* (CHEBI: 67197) has only one *is a* relationship to *cannabinoid* (CHEBI: 67194), while there is a new *is a* to *lipid* (CHEBI: 18059) in the latest (June 2016) version. To date, the errors we found in this study have been submitted to ChEBI via GitHub.

The role of ChEBI in chemistry applications is significant. Therefore, findings of problems at this high level are of major concern. During the past 20 years, SNOMED CT (in which we have seen lower error rates in random concept samples) has been managed by a variety of large professional organizations, such as the College of American Pathologists (CAP), the National Library of Medicine (NLM), and the IHTSDO. Unfortunately, ChEBI does not have the same level of resources that have been available for the maintenance of SNOMED CT. Hence a creative solution to handle the QA chores of ChEBI is needed utilizing ChEBI's curatorial board. It certainly makes sense, as a start, for the curatorial board of ChEBI to conduct a follow-up study of an even larger sample of concepts than the one used in the present study to further assess the error rates in ChEBI for errors of omission and commission.

In future work, ChEBI's curatorial board may want to identify criteria that can be used in locating subsets of concepts that are more likely to be erroneous. As noted, the number of parents as such a criterion did not prove useful, but maybe there are others that will. Employing useful methodologies to help automate aspects of QA efforts should increase the yield of corrections with respect to the curators' expended time. Another potential way to measure the complexity of concepts is by their number of relationships. In a future study, we will test to see if ChEBI concepts with larger numbers of relationships have higher error rates than concepts with fewer relationships. In a

recent study, this was shown to be true with statistical significance for the concepts in the Biological Process hierarchy of the National Cancer Institute thesaurus (NCIt) [19].

V. Conclusions

In this paper, we reported on a quality-assurance (QA) study that was carried out on a sample of ChEBI concepts by two chemistry subject-domain experts. The results revealed that quite a few ChEBI concepts suffer from some kinds of modeling problems. Our consensus report found that 15.5% of the concepts from our sample exhibited severe errors of commission. Particularly prevalent were errors of the type "incorrect and missing chemical classification" and "incorrect charge differences between conjugate acids and conjugate bases." These findings are particularly troubling taking into account the importance of ChEBI and the many applications dependent on it. In general, it appears that the QA processes in place for ChEBI could use further refinement. For example, a targeted effort to review all charge differences between conjugate acids and conjugate bases in ChEBI seems warranted.

Acknowledgment

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health.

References

[1] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016; 44(D1):D1214-1219.

[2] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2008; 36(Database issue):D344-350.

[3] C. UniProt. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43(Database issue):D204-212.

[4] I. Thiele, N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013; 31(5):419-425.

[5] A. Lamurias, J. D. Ferreira, F. M. Couto. Improving chemical entity recognition through h-index based semantic similarity. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S13.

[6] Y. F. Gao, L. Chen, G. H. Huang, T. Zhang, K. Y. Feng, H. P. Li, et al. Prediction of drugs target groups based on ChEBI ontology. Biomed Res Int. 2013; 2013:132724.

[7] D. P. Hill, N. Adams, M. Bada, C. Batchelor, T. Z. Berardini, H. Dietze, et al. Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics. 2013; 14:513.

[8] V. J. Easton, J. H. McCol. Statistics Glossary v1.1, chapter Sampling. 1997.

[9] A. B. Sunter. List Sequential Sampling with Equal or Unequal Probabilities without Replacement. Applied Statistics. 1977; 26(3):261-268.

[10] ChEBI user manual. Available from: http://www.ebi.ac.uk/chebi/userManualForward.do [accessed 4/22/2016].

[11] J. M. Tanur. Margin of Error. In: International Encyclopedia of Statistical Science. Springer Berlin Heidelberg; 2011: 765.

[12] SNOMED CT. Available from: https://www.nlm.nih.gov/healthit/snomedct/ [accessed 4/22/2016].

[13] M. Halper, Y. Wang, H. Min, Y. Chen, G. Hripcsak, Y. Perl, et al. Analysis of error concentrations in SNOMED. Proceedings of the AMIA 2007 Annual Symposium. 2007:314-318.

[14] Y. Wang, M. Halper, D. Wei, H. Gu, Y. Perl, J. Xu, et al. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. J Biomed Inform. 2012; 45(1):1-14.

[15] C. Ochs, Y. Perl, J. Geller, M. Halper, H. Gu, Y. Chen, et al. Scalability of abstraction-network-based quality assurance to large SNOMED hierarchies. Proceedings of the AMIA 2013 Annual Symposium. 2013:1071-1080.

[16] C. Ochs, J. Geller, Y. Perl, Y. Chen, J. Xu, H. Min, et al. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. J Am Med Inform Assoc. 2015; 22(3):507-518.

[17] A. Agrawal, Z. He, Y. Perl, D. Wei, M. Halper, G. Elhanan, et al. The readiness of SNOMED problem list concepts for meaningful use of electronic health records. Artif Intell Med. 2013; 58(2):73-80.

[18] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013; 41(Database issue):D456-463.

[19] S. de Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, L. W. Wright. NCI Thesaurus: using science-based terminology to integrate cancer research results. Stud Health Technol Inform. 2004; 107(Pt 1):33-7.