Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

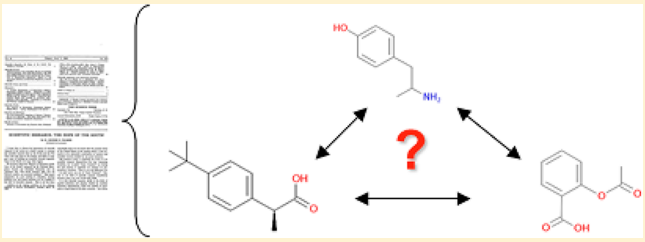
Estimating Error Rates in Bioactivity Databases

Pekka Tiikkainen,*[†] Louisa Bellis,[‡] Yvonne Light,[‡] and Lutz Franke[†]

[†]Merz Pharmaceuticals GmbH, Eckenheimer Landstrasse 100, 60318 Frankfurt am Main, Germany

[‡]EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K.

S *Supporting Information*



ABSTRACT: Bioactivity databases are routinely used in drug discovery to look-up and, using prediction tools, to predict potential targets for small molecules. These databases are typically manually curated from patents and scientific articles. Apart from errors in the source document, the human factor can cause errors during the extraction process. These errors can lead to wrong decisions in the early drug discovery process. In the current work, we have compared bioactivity data from three large databases (ChEMBL, Liceptor, and WOMBAT) who have curated data from the same source documents. As a result, we are able to report error rate estimates for individual activity parameters and individual bioactivity databases. Small molecule structures have the greatest estimated error rate followed by target, activity value, and activity type. This order is also reflected in supplier-specific error rate estimates. The results are also useful in identifying data points for recuration. We hope the results will lead to a more widespread awareness among scientists on the frequencies and types of errors in bioactivity data.

INTRODUCTION

Chemical and biological databases are vital for drug discovery and development. Large quantities of genomic and proteomic data have been publicly available for years thanks to a revolution in genome sequencing. In contrast, the field of small molecule activity databases was until few years ago the realm of commercial suppliers. Curating bioactivity data from articles and patents is laborious and therefore expensive, meaning that mainly pharmaceutical companies and the like with resources to pay the high annual fees have had access to these databases.

The explosion in publicly available bioactivity data[1] has spurred a wave of activity in the research community analyzing the ocean of data. Examples of this include the works on binding affinity variability caused by assay setup[2] and target protein source species.[3] Complementarity of public and commercial bioactivity databases has been studied by both Southan et al.[4,5] and Tiikkainen and Franke.[6]

Quality of these databases is crucial for everyone in the cheminformatics and molecular modeling communities. Incorrect ligand structures or activity values in training data sets can lead to misleading predictions and research decisions. Recently the field has been paying more attention to errors in commonly used databases[7,8] with efforts such as MIABE[9] being initiated to create industrywide standards for bioactivity data reporting and storage.

In our previous work,[6] we performed a detailed analysis of content and data overlap of several bioactivity databases. We also reported inconsistencies in data extracted from the same source documents by different data suppliers. In the current work, we have analyzed these inconsistencies in more detail by comparing three large bioactivity databases (WOMBAT,[10] ChEMBL,[11,12] and Liceptor[13]). As a result, we have derived parameter-specific error rate estimates for each of the three databases.

MATERIALS AND METHODS

Data Resource. Bioactivity data used in the present work are all drawn from an internaldata resource called the MVBD (Merz Virtual Bioactivity Database), which integrates bioactivity data from a number of public and commercial data suppliers. For this publication, a bioactivity is defined as a combination of a protein target (given as a Uniprot accession number[14]), ligand structure, activity value (given as micromolar), activity relation (e.g., ">" denotes that potencies were greater than the activity value), and activity type (e.g., IC_{50} or K_i).

Overview of data sources used in this study is given in Table 1. Overlap of the data sources in terms of bioactivities is shown in Figure 1. The importance of not relying on a single supplier is highlighted by the fact that 92.5% of the bioactivities are available from one supplier only. ChEMBL and Liceptor have the highest proportion of unique bioactivities, which is explained, by their heterogeneous data sources. In addition to citing journal articles, ChEMBL contains large data sets from Pubchem[15] and various screening campaigns while Liceptor cites more than 30 000 patents. For a more detailed analysis of the data resource, the reader is referred to our earlier paper on the subject.[6]

Received: February 11, 2013
Published: October 2, 2013

Table 1. Key Statistics for the Three Databases Considered in the Analysis

database	version	journal articles	patents	bioactivities[a]	ligands[b]	targets[c]
ChEMBL	release 14	45144	0	3296302	732189	2755
WOMBAT	2012.01	15601	0	394051	187964	1486
Liceptor	2012_03	12751	31801	1769705	846877	860

[a] Number of unique standardized bioactivities. A bioactivity is a combination of target, ligand, activity value, activity relation, activity unit, and activity type. [b] Ligands which are involved in at least one standardized bioactivity. [c] Targets which are involved in at least one standardized bioactivity. Orthologs of the same protein were considered as a single target. Please note that all these figures are a filtered interpretation of the three databases, and they do not include bioactivities with a nonprotein target (e.g., a cell or an organism) or a nonquantitative activity value. Therefore the figures quoted by the suppliers themselves can greatly differ from the numbers in this table.

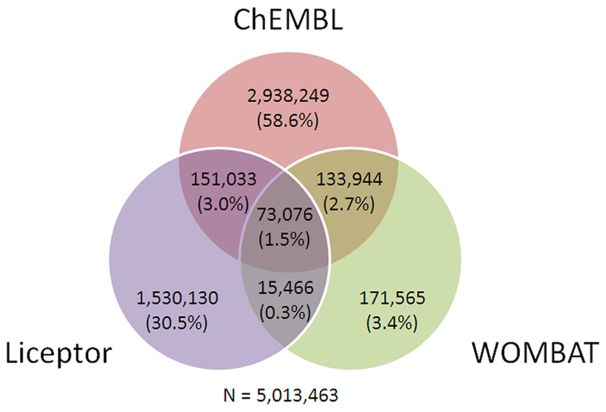


Figure 1. Overlap of bioactivities in ChEMBL, Liceptor, and WOMBAT. Figures in parentheses denote the share of all bioactivities within the sector.

All bioactivity parameters have been standardized to allow a direct comparison of data points. Target identity is usually given by the suppliers as Uniprot accessions[14] making integration straightforward. In some cases, an internal identifier has been used by a supplier which had to be converted into a Uniprot accession using a look-up table. Activity values can be either linear or logarithmic with a variety of concentration units attached (micromolar and nanomolar being the most popular). These have been standardized into linear micromolar values. Standardization of chemical structures requires the most attention. Our procedure first removes any known counterions and salts from the structure, followed by calculation of the canonical tautomer and deprotonation of bases and protonation of acids. Finally stereochemistry is standardized and only the largest fragment remaining in the structure is retained. All these steps have been implemented as a workflow in Pipeline Pilot 8.5.[16] To allow backtracking to the original data source, the original nonstandardized activity parameter values are also stored in the MVBD. Mapping of supplier field names to bioactivity parameters is given in Table 2.

Error Rate Estimation. Before describing the approach taken, we would like to clarify two key terms used below: individual cases where suppliers disagree on a parameter value are called discrepancies while statistics calculated on discrepancies are called error rate estimates.

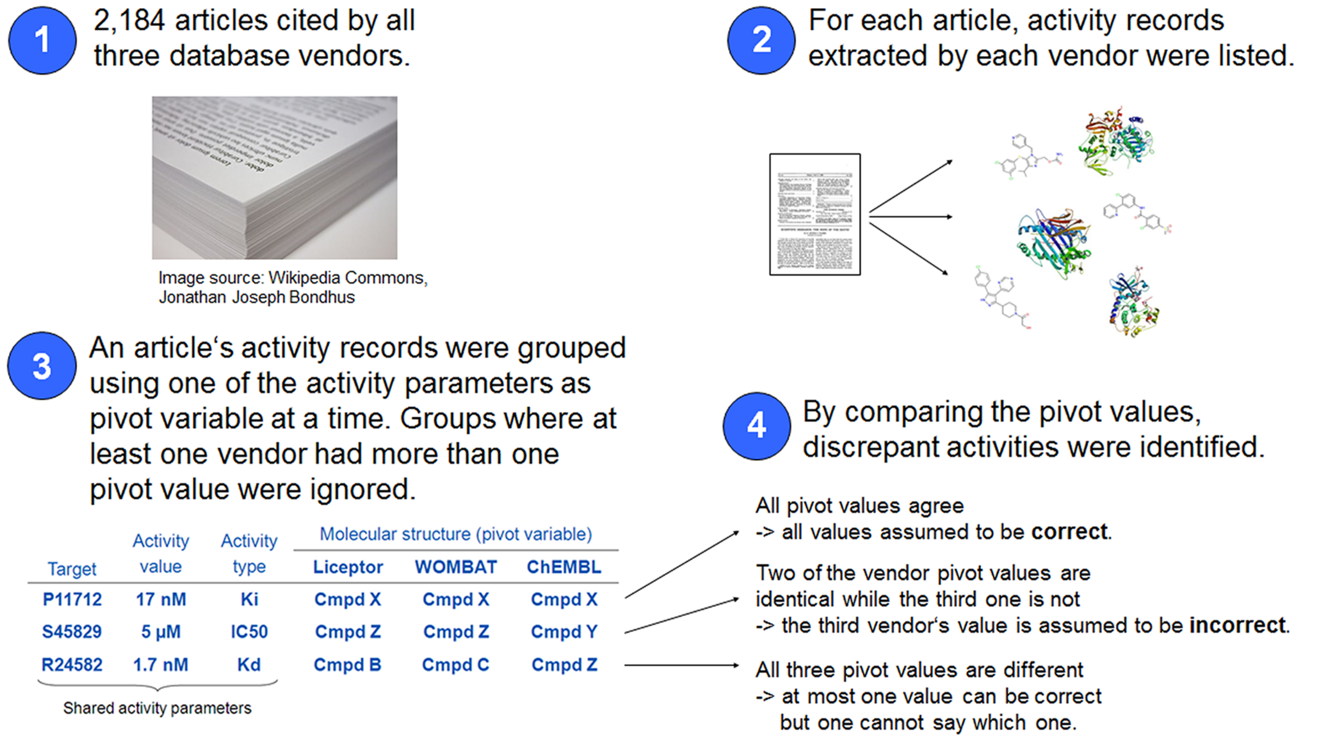
The logic of our approach for estimating error rates lies on the assumption that given the same source article, each database supplier's curation effort would result in the same bioactivity data being saved into their databases. Any discrepancies between suppliers in data extracted from the same article would imply that mistakes have been made in the curation step. Since our integrated bioactivity database includes data from three large data suppliers (ChEMBL, Liceptor, and WOMBAT), we were able to identify and quantify these discrepancies on a large scale resulting in error rate estimates for the curation step.

Error rate estimates were calculated by comparing bioactivity records that the three suppliers had extracted from the same source article (Figure 2), with 2184 articles being cited by all three. Independently for each article, the bioactivities were pivoted into groups where all except the pivot variable had to contain identical values for all suppliers. If the value of the pivot variable agreed for all three suppliers, all bioactivities in the group were deemed to be correct. If two suppliers had the same pivot variable value but the third one disagreed, the latter supplier's value was deemed incorrect. If all three suppliers' pivot variable values were in disagreement, no conclusions were drawn. It should be noted that only bioactivity groups with exact activity values (denoted as an equality sign (=) in the relation field) were included in the analysis. Additionally, if any of the suppliers had more than one value in the pivot variable,

Table 2. Names of the Original Supplier Fields from Which the Central Bioactivity Parameters Were Read

activity parameter	ChEMBL[a]	Liceptor[b]	WOMBAT
original data format	MySQL database	RD file	RD file
ligand structure	compound_structures. canonical_smiles	structure entry from the RD file	structure entry from the RD file
target	target_dictionary. protein_accession	target entry from the RD file.\	ROOT:ACT.LIST(): SWISSP.ID
activity value	activities.standard_value	value entry from the RD file	ROOT:ACT.LIST():value
activity unit	activities.standard_units	unit entry from the RD file	all activity values are logarithmic and therefore no activity unit is needed
activity type	activities.standard_type	activity type entry from the RD file	ROOT:ACT.LIST():type
activity relation	activities.relation	relation entry from the RD file	parsed from the activity value

[a] ChEMBL was sourced from a MySQL database, and the cell contents give the database table and field names, separated with commas. [b] The supplier of Liceptor considers the field names to be confidential information, and therefore, the exact field names are not given here.



1 2,184 articles cited by all three database vendors.

Image source: Wikipedia Commons, Jonathan Joseph Bondhus

2 For each article, activity records extracted by each vendor were listed.

3 An article‘s activity records were grouped using one of the activity parameters as pivot variable at a time. Groups where at least one vendor had more than one pivot value were ignored.

Target	Activity value	Activity type	Molecular structure (pivot variable)		
			Liceptor	WOMBAT	ChEMBL
P11712	17 nM	Ki	Cmpd X	Cmpd X	Cmpd X
S45829	5 µM	IC50	Cmpd Z	Cmpd Z	Cmpd Y
R24582	1.7 nM	Kd	Cmpd B	Cmpd C	Cmpd Z

Shared activity parameters

4 By comparing the pivot values, discrepant activities were identified.

All pivot values agree
-> all values assumed to be **correct**.

Two of the vendor pivot values are identical while the third one is not
-> the third vendor‘s value is assumed to be **incorrect**.

All three pivot values are different
-> at most one value can be correct
but one cannot say which one.

Figure 2. Approach for estimating error rates. Bioactivities extracted from the same article by all three vendors were pivoted, and by comparing the pivot values, the data points were declared to be correct or incorrect.

the respective bioactivity group was ignored. These filters were used to reduce ambiguity in the analysis.

This comparison was repeated for all four bioactivity variables considered in this work (ligand structure, protein target, activity value, and activity type) by using each of these as the pivot variable. Summing up types of discrepancies for each activity parameter resulted in error rate estimates for each data supplier and parameter.

The number of bioactivities extracted from the 2184 common articles varied for each supplier, with ChEMBL containing 125 660 activities, followed by WOMBAT’s 102 109 entries and Liceptor’s 72 472. ChEMBL reports bioactivity data of protein complex targets by splitting the entry for each subunit while WOMBAT and Liceptor usually assign only one of the subunits as the target. This duplication explains in part ChEMBL’s large bioactivity count. It should be noted that this analysis of data was carried out on ChEMBL version 14 and that handling of protein subunits has changed from version 16 onward, with the target protein complex given as one target (Mark Davies, ChEMBL Group, personal communication).

After the pivoting step, there were 41 806 activity groups where the ligand structure was used as the pivot variable. The corresponding group counts for the target, activity value, and activity type parameters were 44 418, 42 431, and 43 840, respectively.

It should be pointed out that the approach used cannot identify errors in the original source document which might have arisen, for example, due to imperfect assay conditions or typing errors by the source document authors. Therefore, the aggregate error rate taking into account all kinds of error sources is likely to be greater than the figures reported below. Readers interested in additional sources of errors are referred to a review by Kramer and Lewis.[17]

Ligand Structure Discrepancies. Ligand structure discrepancies can be further divided into those where the atom connectivity differs and those where the connectivity is shared but the stereochemistry is not accounted for or a discrepant isomer is given. The former discrepancy type is more serious affecting all applications of the data whereas the latter is detrimental mainly to tools requiring three-dimensional models of the ligands.

In order to differentiate between the two types of ligand structure discrepancies, we took all pivoted activity groups where at least one supplier had a discrepant ligand structure. The first 14 characters of the ligands’ InChI keys[18] encode the atom connectivity while the remaining characters describe stereochemical information. If these first 14 characters differed across suppliers, the error was deemed to be of the first type (incorrect atom connectivity). Otherwise, the discrepancy was due to the stereochemical presentation but not to the atom connectivity.

Target Identity Discrepancies. Analogous to ligand structures, discrepancies in protein targets can also be split into two categories. The first category contains those where a discrepant ortholog has been assigned to the protein (e.g., rat protein is given as the target instead of the human ortholog). In the second category, the target itself is discrepant. In our work, we sought to quantify the frequency of the two discrepancy types. Activity groups were identified where the protein target (given as a species-specific Uniprot accession[14]) was different for at least one supplier. The Uniprot accessions were converted into the corresponding recommended protein names[19] shared by all orthologs. If the recommended protein names were the same for the all suppliers, the discrepancy was deemed to be of the first type (target shared but species different) and of the second type otherwise.

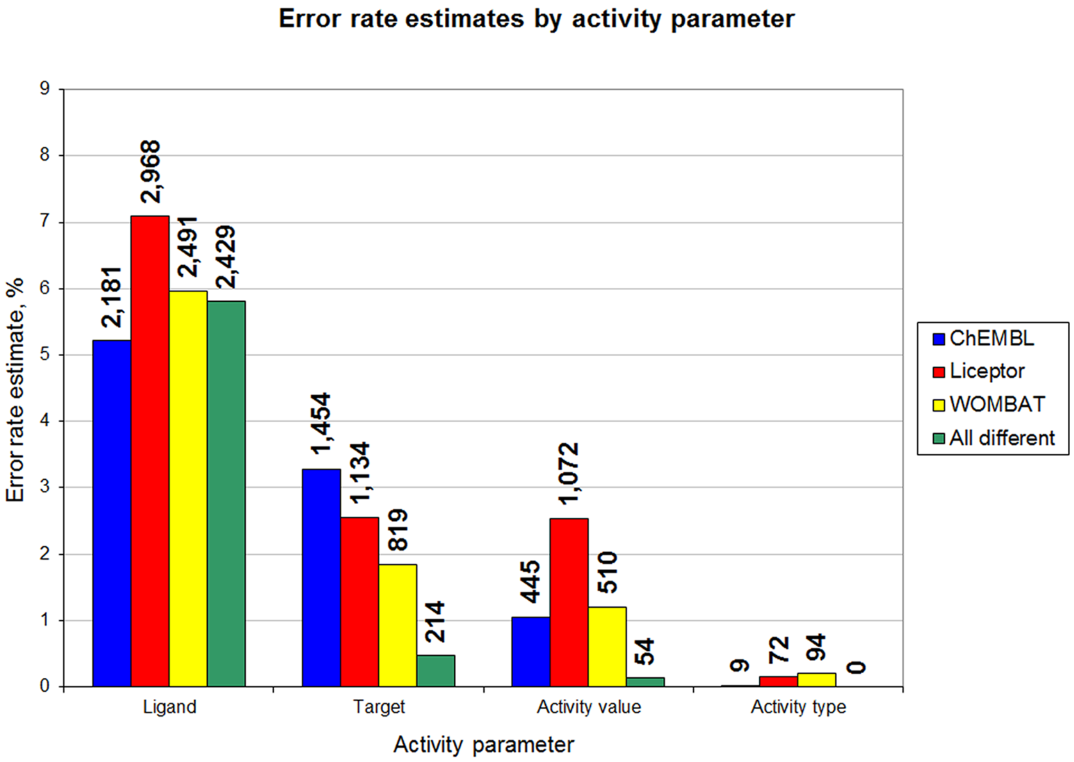


Figure 3. Activity parameter specific error rate estimates for each data supplier plus frequency of cases where all three suppliers had differing values. Figures above the bars indicate the absolute number of activities.

Method Validity. The implicit assumption behind the error rate estimation method described above is that all three suppliers have curated the data independent of each other. If at least two suppliers had recorded identical bioactivity records, these were assumed to be correct as it is unlikely that exactly the same mistakes would have been made in the curation process.

To verify this hypothesis, we manually inspected randomly selected groups of bioactivity records where one supplier's pivot variable value differed from that of the other two suppliers and therefore was deemed incorrect. Forty-five such cases were checked: five cases per supplier for three activity parameters (ligand structure, target and activity value). A check was not made for the activity type parameter due to its very low discrepancy frequency. For each case, we compared the values to the one reported in the original paper and recorded if the discrepant value really was incorrect as assumed. Only activities extracted from the ACS's *Journal of Medicinal Chemistry* articles were checked due to journal access reasons. Detailed results are given in Supplementary Tables S1a−c.

■ RESULTS AND DISCUSSION

Validity of the Approach. To quantify the reliability of the error rate estimates given below, we manually checked 45 potentially erroneous activity data points identified by our approach (Supporting Information Tables S1a−c). In 37 cases (82.2%) the discrepant activity parameter value in fact turned out to be incorrect. In three cases (6.7%), the opposite was true (the two suppliers with agreeing values were in fact wrong). In five cases (11.1%) we were not able to rule an activity wrong or correct because of ambiguity and/or lack of clarity in the source article (e.g., cell lines from different species used but authors fail to specify which data point was measured with which cell line).

Additional validation comes from the actual recuration of most of the discrepancies identified in the ChEMBL database. There were 1936 unique ligands among the 2181 bioactivities where ChEMBL was the only one with a discrepant ligand. Here, 310 of these ligands (16.0%) turned out to have the correct structure. For the remaining 1626 ligands some changes had to be made, which ranged from a complete redraw of the structure to very simple changes such as the addition of a stereo bond or a salt to the compound. Also the 2429 bioactivities where all suppliers disagreed on the ligand structure were checked. Out of the 1486 ChEMBL ligands in this set, 280 (18.8%) were in agreement with the source article and therefore correctly curated leaving 1206 ligands for which some changes were necessary. Out of these remaining 1206 ligands, 761 (51.2%) only needed the addition of a stereo bond to the structure; or, the stereochemistry that was shown in the paper could not be accurately represented in the molfile, and so the structure could not be changed. In 101 compounds (6.80%), the stereochemistry was incorrect and needed to be changed. In the cases where either the activity type or value was discrepant in the ChEMBL, 259 have been corrected in ChEMBL release 16. The ChEMBL bioactivity value was found to be correct in 83 cases (32.0%) meaning that our approach identified an erroneous value correctly in 68.0% of the cases. Likewise 764 bioactivities (corresponding to 88 assays) were checked where ChEMBL had a discrepant target. For 137 of these bioactivities (37 assays) the target assigned was found to be correct (18.0%). For the remaining 627 bioactivities (corresponding to 51 assays, 82.0%) either the target (21 bioactivities, 2.7%) or the ortholog (606 bioactivities, 79.3%) have been corrected. As in the small-scale validation above, the ChEMBL team came across original papers where a bioactivity

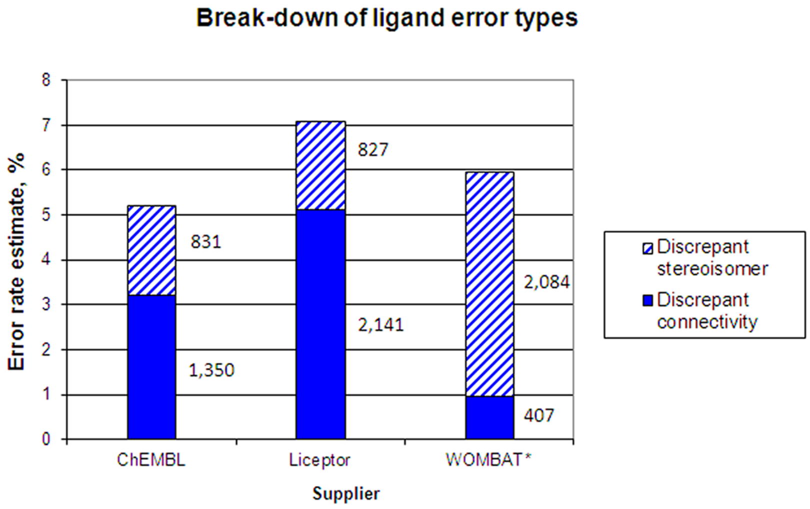


Figure 4. Breakdown of types of discrepancy errors. The solid bars give the frequency of discrepant atom connectivities while the remaining part is due to discrepant or missing stereochemistry. The figures next to the bars denote bioactivity counts. * A large share of WOMBAT's discrepant stereoisomers are probably not erroneous but rather due to the other two suppliers ignoring stereochemistry in some of the bioactivities under comparison. Please refer to the text for details.

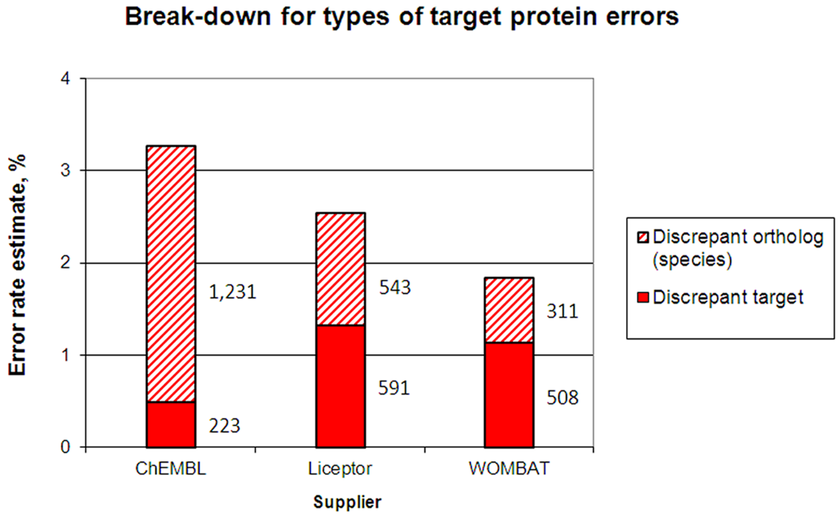


Figure 5. Error rate estimates for target protein identity. These can be divided into two categories: discrepancies where the target is correct but the source species (ortholog) is discrepant and those where the target is discrepant. The figures next to the bars are absolute bioactivity counts.

parameter had not been clearly reported making accurate curation impossible.

Since 68–84% of the discrepancies checked have required some changes, we can argue that the approach we have used is fairly accurate in identifying errors.

Error Rate Estimates. Figure 3 gives supplier-specific error rate estimates for the four activity parameters considered in the current work (ligand structure, activity value, activity type, and protein target). Ligand structures have the highest probability of being discrepant followed by the protein target, activity value, and finally the activity type. The relative order of the estimated error rates reflects the complexity of the respective parameter. A ligand structure has dozens of features (bonds, atoms, and types thereof) that lead to more possibilities for curation errors, hence the high error rate observed. Errors in activity values are mainly due to unit conversion issues (e.g., micromolar affinities curated as nanomolar). Finally, the activity type has only a handful of options to choose from (IC_{50}, K_i, etc.), and these are usually clearly stated in the source articles.

Frequencies for cases where all three suppliers report different values (green bars in Figure 3) are relatively low for all parameter types except for the ligand structure. In these cases at most only one of the suppliers can have the correct value. If these cases were manually checked from source articles and the identified errors assigned to the suppliers, the supplier-specific error rates would be higher than those reported in Figure 3.

Breakdown of types of ligand discrepancies is shown in Figure 4. In Liceptor and ChEMBL, differences in stereochemistry account for 30–40% of all ligand structure discrepancies with the rest being discrepancies in the atom connectivity. In WOMBAT, some 80% of ligand structure discrepancies are due to stereochemistry. It should also be noted that sometimes the discrepancy arises because the discrepant supplier is the only one with any stereo atoms or bonds defined for the ligand: the frequency is 51% of stereochemistry discrepancies for WOMBAT, 6.0% for ChEMBL, and 6.5% for Liceptor. This finding led us to suspect whether the lone supplier providing the stereoisomer in fact would be correct, with the other two suppliers simply omitting stereochemistry in some of the ligands under comparison for discrepancies. We checked ten such cases for WOMBAT and found out that in eight of these cases a stereoisomer had actually been defined in the source document and was correctly recorded in WOMBAT while one case was ambiguous (Supporting Information Table S1d). Therefore, the true error rate for WOMBAT ligands is likely to be lower than the estimate given in Figure 4.

As shown in Figure 5, most of the discrepancies in target protein identity are due to a different ortholog (source species) assigned for the correct target. ChEMBL has the highest target error rate estimate of 3.27%, but 85% of the discrepancies are due to a discrepant species assigned leaving only 0.5% of all ChEMBL activities assigned a discrepant target. This observation is partially explained by the fact that in many of these cases, the true target is a guinea pig, rabbit, or hamster protein for which no sequence had been available at the time of curation. Instead, the human ortholog, for which a sequence existed, had been assigned as the target. Analogous to incorrect stereoisomers in ligands, having the wrong species but the correct target is less serious than having also the target wrong. Nevertheless, this could be an issue if one is interested in the species-specificity of a ligand.

Recuration of Discrepant Values. To improve the quality of the bioactivity databases studied, we have sent each supplier a list of discrepant bioactivities identified in their respective databases. Using the list we provided, the curation team of the ChEMBL database, as described above, has already corrected a large number of activities and ligands. All the changes required will be incorporated into ChEMBL release 17. According to the company representatives for the two commercial databases, also the discrepancies identified in WOMBAT (Tudor Oprea, Sunset Molecular Discovery Ltd., personal communication) and Liceptor (Aniket Ausekar, Evolvus Group, personal communication) are going to be reviewed.

CONCLUSIONS

In the current work, we have given estimates for error rates in bioactivity databases arising from the curation step of data retrieval, i.e. errors introduced when bioactivity data in an article is converted into a digital format. The method also has value in highlighting bioactivity data that require recuration by the database suppliers.

The approach taken is simple but robust and less laborious than manually checking thousands of activity data points from each database and deriving the error rate estimates from these. However, it must be kept in mind that the underlying assumption, i.e. an activity parameter value is correctly curated if at least two databases agree on it, does not always hold. The exceptions inflate the error rate estimate for the discrepant supplier while lowering the estimate for the other two suppliers. These exceptions to the rule can have several reasons, e.g. data exchange between databases we are not aware of, where incorrectly curated data gets replicated into another database, and more careful curation effort by one of the suppliers as seen with extraction of stereochemical information by WOMBAT. Fortunately these exceptions are outnumbered by cases where the approach correctly identifies an error—as shown by our small-scale internal validation and the recuration work done within the ChEMBL group. Additionally, all suppliers are affected by these exceptions, leading to cases canceling each other out further reducing the effect on error rate estimates.

Benefits to the research community arising from our work are threefold. First, we hope that scientists keep in mind the risk of errors when using data from bioactivity databases. Those training machine learners and other QSAR models with data from bioactivity databases are encouraged, if possible, to double-check their data points from the original source documents. This should help to maximize the quality of their models.[17,20] Second, the individual bioactivities flagged as potentially erroneous by our method are being recurated and corrected by the database suppliers, which will directly improve the quality of the databases. Third, the results should provide a guideline for improving the curation process in the future by drawing closer attention to the correctness of ligand structures and protein targets, the two parameters with the greatest estimated error rates.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Tables S1a−d contain results of manual method validation. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pekkatii@luukku.com.

Author Contributions

P.T. devised the study and performed the discrepancy analysis and error rate estimation. L.B. and Y.L. recurated the discrepant ChEMBL bioactivities and ligands.

Notes

The authors declare the following competing financial interest(s): L.B. and Y.L. are employees of the European Bioinformatics Institute which maintains one of the databases (ChEMBL) analyzed in this work.

ACKNOWLEDGMENTS

Dr. Yusuf Tanrikulu is credited for reading through the manuscript and giving valuable comments.

REFERENCES

(1) Bender, A. Databases: Compound bioactivities go public. *Nat. Chem. Biol.* **2010**, *6*, 309−309.

(2) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The experimental uncertainty of heterogeneous public K(i) data. *J. Med. Chem.* **2012**, *55*, 5165−5173.

(3) Kruger, F. A.; Overington, J. P. Global analysis of small molecule binding to related protein targets. *PLoS Comput. Biol.* **2012**, *8*, No. e1002333, http://10.0.5.91/journal.pcbi.1002333 (accessed January 10, 2013).

(4) Southan, C.; Varkonyi, P.; Muresan, S. Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. *Curr. Top. Med. Chem.* **2007**, *7*, 1502−1508.

(5) Southan, C.; Varkonyi, P.; Muresan, S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminf.* **2009**, *1*, No. 10, http://www.jcheminf.com/content/1/1/10 (accessed Feb 24, 2013).

(6) Tiikkainen, P.; Franke, L. Analysis of commercial and public bioactivity databases. *J. Chem. Inf. Model.* **2012**, *52*, 319−326.

(7) Williams, A. J.; Ekins, S. A quality alert and call for improved curation of public chemistry databases. *Drug Discovery Today* **2011**, *16*, 747−750.

(8) Williams, A. J.; Ekins, S.; Tkachenko, V. Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discovery Today* **2012**, *17*, 685−701.

(9) Orchard, S.; Al-Lazikani, B.; Bryant, S.; Clark, D.; Calder, E.; Dix, I.; Engkvist, O.; Forster, M.; Gaulton, A.; Gilson, M.; Glen, R.; Grigorov, M.; Hammond-Kosack, K.; Harland, L.; Hopkins, A.; Larminie, C.; Lynch, N.; Mann, R. K.; Murray-Rust, P.; Lo Piparo, E.; Southan, C.; Steinbeck, C.; Wishart, D.; Hermjakob, H.; Overington, J.; Thornton, J. Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discovery* **2011**, *10*, 661−669.

(10) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*, 1st ed.; Oprea, T. I., Ed. Wiley: New York, NY, 2005; Vol. *23*, pp 223−239.

(11) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.

(12) European Bioinformatics Institute. *ChEMBLdb*. https://www.ebi.ac.uk/chembldb/index.php (accessed December 15, 2012).

(13) Evolvus. *Liceptor database*. http://www.evolvus.com/Products/Databases/LiceptorDatabase.html (accessed January 17, 2013).

(14) Uniprot Consortium. Uniprot accession numbers. http://www.uniprot.org/manual/accession_numbers (accessed June 15, 2012).

(15) National Center for Biotechnology Information. The PubChem Project. http://pubchem.ncbi.nlm.nih.gov/ (accessed May 6, 2013).

(16) *Pipeline Pilot*, version 8.5; Accelrys, Inc.: San Diego, CA, 2011.

(17) Kramer, C.; Lewis, R. QSARs, Data and Error in the Modern Age of Drug Discovery. *Curr. Top. Med. Chem.* **2012**, *12*, 1896−1902.

(18) International Union of Pure and Applied Chemistry. The IUPAC International Chemical Identifier. http://www.iupac.org/inchi (accessed August 10, 2012).

(19) Uniprot Consortium. Protein naming guidelines. http://www.uniprot.org/docs/nameprot (accessed Nov 27, 2012).

(20) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.