www.csdata.org

ISSN 2096-2223
CN 11-6035/N

文献 DOI:
DOI: 10.11922/csdata.2018.0061.zh
数据 DOI：
DOI: 10.11922/sciencedb.657
文献分类：化学

收稿日期：2018-10-08
开放同评：2018-10-12
录用日期：2019-02-18
发表日期：2019-03-22

化学药物数据集

徐挺军[1*]，赵英莉[1]，李英勇[1]

1. 中国科学院上海有机化学研究所信息中心，上海 200032

摘要：通过收集书籍手册中的化学药物研究成果信息，对收集的药物数据信息进行分类汇总和规范化处理，采用化合物唯一标识法集成药物化合物结构数据，利用数据采集规范和数据抽查回溯手段控制数据质量，通过算法程序保证数据集中关键数据项的正确率，最终形成化学药物数据集。本数据集包括了药物基础数据和药物化合物数据，共计 1060 条。化学药物数据集可以为新药研发、药物改良、科研教育等提供数据支持。

关键词：药物数据；化学药物；药物化合物；新药研发

数据库（集）基本信息简介

数据库（集）名称	化学药物数据集
数据作者	徐挺军、赵英莉、李英勇
数据通信作者	徐挺军（xutingjun@sioc.ac.cn）
数据时间范围	1985–2001年
地理区域	世界各国
数据量	5.44 MB
数据格式	*.MDB
数据服务系统网址	http://www.sciencedb.cn/dataSet/handle/657
基金项目	中国科学院信息化专项科学大数据工程（XXH135）、上海市化学化工数据共享服务平台（18DZ2294000）
数据库（集）组成	数据集由2部分数据组成：1. 药物基础数据（包括药物类型、名称、性状、制法、用途、生产企业等）；2. 药物化合物数据（包括化合物登录号、CA登记号、化合物名称、分子式、分子量、化学结构文件等）。

引 言

化学药物是当今世界占比最高的药物来源，其数量众多、研发活跃、发展迅速。但是，我国化学制药行业严重缺乏竞争能力，相比于其他发达国家的医药产业，具有技术创新能力低、研发投入少、仿制药物占比高等弱点[1]。药物研发成本高、周期长、技术保护等因素制约着我国合成药物的创新和发展，如何能准确地找出突破点和应对方法是整个医药产业链值得深思的问题。

药物化学家通过研究现有药物化合物明确的靶标结构和物性活性数据，基于化学原理，根据药物的化学结构特征、合成方法等，构建新的药物化学有效结构

* 论文通信作者
徐挺军：xutingjun@sioc.ac.cn

类型或者新的药物合成路径，进行药物模拟创新，成为突破现阶段我国药物创新困境和瓶颈的一个方法[2]。对于原创新药研发投入高、失败率高等问题，研究现有药物的构效关系，发现现有药物新的用途或者新的定位，能够在一定程度上提升新药研发的成功率，降低药物研发成本，加快临床急需用药的上市[3]。

通过收集国内外已经上市的药物以及有发展潜力、尚在研发中的新药等现有药物的基础数据和药物化合物的数据，对数据进行加工和规范化处理，形成一定规模和范围内的化学药物数据集，涵盖药物的物性活性、制法合成路径、化合物结构等信息数据，从而从数据的角度促进药物研发等科研活动的进行。

国内《化学专业数据库》中的药品数据库，收集了约 9000 多种药品，数据包括药品的名称、结构、理化性质、适应症、标准等，由于是多种数据源的整合，存在数据规范不统一、药物分类不清晰等问题，且没有药物制法、生产企业等数据[4]。世界著名的药品数据库 DrugBank，最新版本涵盖了约 10000 多种药物，其中化学小分子药物 2000 余种，主要为药物药理学数据和药物靶点数据，数据描述语言为英语[5]。本文希望通过化学药物数据集的建设，以小范围的典型数据源为例，研究化学药物数据的采集、处理方法和步骤，为建立更加权威、规范、全面的中文化学药品数据库打下基础。

1 数据采集和处理方法

1.1 原始数据来源

化学药物数据集的原始数据采集自《精细化工产品手册·药物》[6]，原始数据采集后对数据进行规范化加工处理，然后对药物化合物进行唯一化标识[7]，获得药物化合物的结构信息数据，最终形成化学药物数据集。

1.2 数据采集

原始数据为手册书籍，其编写按用途、药理和化学结构相结合的方式进行分类。对于有共同药理作用的药物，如拟肾上腺素和抗肾上腺素药物、拟胆碱和抗胆碱药物、抗组胺药物等，分列一章。每章开头有对该章药物的简短说明。每章中再分小类，在小类中将结构相似的药物归于一起。同一种药物有多种用途时，该药物归在主要用途一章中。在药物信息详细描述段中，均分栏介绍其中文通用名（或常用名）及英文通用名（或常用名）、在美国《化学文摘》上的登录号、其他名称、结构式、分子式、相对分子质量、性状、制法、规格、用途、生产厂家、参考资料等，如图 1 所示。

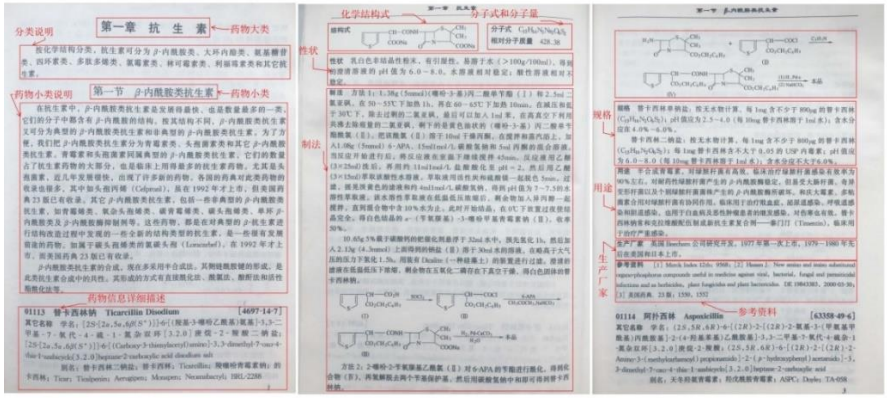


图 1 原始数据示例

根据以上原始数据类型和特点，设计数据集录入加工数据表，采用人工输入的方式，将原始数据书籍中各个信息分别录入对应的数据项中。数据表以化学药物作为实体，药物的分类信息、药物描述数据、化合物数据、性质数据、生产方法等作为其属性。由于原始书籍涉及的数据类型众多，采用一定的输入规则对数据进行采集，以便后续的数据规范化处理，数据部分采集规则如表 1 所示。

表 1　数据采集规则

序号	规则说明
1	同一类数据中有多个数据的使用“；”分隔，如同一个药物有多个名称，多个参考文献等
2	内容描述中每个自然分段的结尾部分加“$$”
3	如果内容中有表格，在表的起始和结束后各加一个“$T”，表内容的每一项用“/”或“@”（当表的内容中有“/”时）分隔，每一行用“$$”分隔
4	化学结构式中的结构图不需输入
5	所有汉字和符号，包括希腊字母按原样输入
6	分子式中的数字按普通数字方式输入，其余上下标内容使用上标符组“^<”“^>”，和下标符组“^{”“^}”表示，需要用上下标表示的内容置于上标或下标符号组的两个符号间。例如，“cm^{-1}”应该表示成“cm^<-1^>”
7	熔点（mp）、沸点(bp)、酸碱度（PH）的数据包含在性状内容中，需要从中选取，输入内容包括这些数据的标识。

1.3　数据规范化处理

原始数据经采集后，形成化学药物加工数据表。其中药物大类为药物的主要用途分类，药物小类为化学结构或者药理作用部位分类，如抗生素大类中，分 β-内酰胺类抗生素、大环内酰胺类抗生素、氨基糖苷类抗生素等小类。根据化学药物数据集的设计，将加工数据表中的数据进行规范化处理：去除 CASRN 号中的“-”，将其转换为数字以便后续的数据处理；设计药物基础数据表、药物化合物数据表，分别如表 2、表 3 所示，并将加工数据表中不同类型的数据分别归类至相应的数据表中，并以药物编号 YWID 作为主键链接；由于药物化合物的化学结构大多较为复杂，如采用人工输入化学结构数据，则需要非常专业的人员耗费相当多的工作时间，且较易出现差错。化学药物数据集利用原始数据中较为明确的 CASRN 号、化合物名称、分子式等数据，采用化合物唯一化标识方法，根据化合物登录号 SRN 直接从化合物参考数据库中获取化合物结构信息[8]，形成药物基础数据、药物化合物数据，得到最终的化学药物数据集。

表 2　药物基础数据表

序号	属性名称	数据类型	属性说明
1	YWID	数值	药物编号
2	YWDL	字符	药物大类
3	YWXL	字符	药物小类
4	YWZW	字符	药物通用名称
5	YWYW	字符	药物通用名称英文

序号	属性名称	数据类型	属性说明
6	XZ	字符	性状
7	RD	数值	熔点
8	FD	数值	沸点
9	PH	数值	酸碱度
10	ZF	字符	制法
11	YT	字符	用途
12	SCCJ	字符	生产企业
13	CKWX	字符	参考文献

表 3　药物化合物数据表

序号	属性名称	数据类型	属性说明
1	ID	数值	序号
2	YWID	数值	药物编号
3	SRN	数值	化合物登录号
4	CASRN	字符	CA 登记号
5	HHWM	字符	化合物名称
6	QTMC	字符	化合物别名
7	MF	字符	分子式
8	MW	数值	分子量
9	Mol	字符	化学结构文件

2　数据样本描述

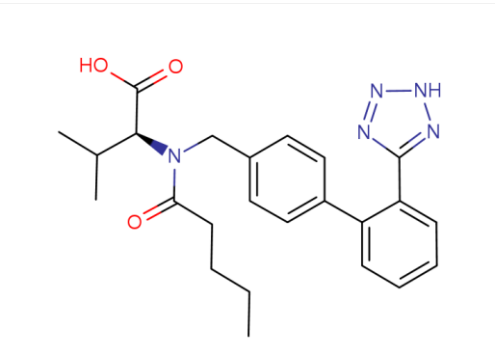
化学药物数据集涵盖了药物的基础信息数据和药物化合物数据，分别存储于药物基础数据表和药物化合物数据表。以市面常见的抗高血压药物缬沙坦（Valsartan）为例，表 4 展示了其药物的基本信息数据，表 5 展示了其化合物数据，其中化学结构数据为 MDL Molfile 文件内容，在表中以化学结构图的形式来描述。

表 4　化学药物数据集药物基础数据示例

序号	数据类型	数据示例
1	药物编号	374
2	药物大类	心脑血管药物
3	药物小类	抗高血压药
4	药物通用名称	缬沙坦
5	药物通用名称英文	Valsartan
6	性状	从二异丙醚结晶，熔点 116–117℃。
7	熔点	116–117℃

序号	数据类型	数据示例
8	沸点	-
9	酸碱度	-
10	制法	2’-氰基联苯-4-醛(I)和 L-缬氨酸甲酯进行还原胺化，得到的化合物(II)再用戊酰氯进行酰化，层析后得到化合物(III)。然后和 Bu3SnN3 进行反应，引入四唑，再水解即得产物。
11	用途	抗高血压药物。非肽血管紧张素 II AT1-受体拮抗剂。用于治疗高血压。
12	生产厂家	瑞士 Ciba 开发，1996 年在德国上市。
13	参考文献	[1] Merck Index 12th：10051；[2] Buehlmayer P，Ostermayer F and Sehmidlln T．Aeyl compounds．EP 443983，1991-08-28；$$[3]Buehlmayer P，Ostermayer F and Schmidlin T．Acyl compounds．US 5399578，1995-03-21.

表 5　化学药物数据集药物化合物数据示例

序号	数据类型	数据示例
1	序号	382
2	药物编号	374
3	化合物登录号	6137969
4	CA 登记号	137862-53-4
5	化合物名称	N-(1-氧戊基)-N-[[2’(-(1H-四唑-5-基)[1,1’(-联苯]-4-基]甲基]-L-缬氨酸；N-(1-Oxopentyl)-N-[[2’-(1H-tetrazol-5-y1)[1,1’-biphenyl]-4-y1]methyl]-L-valine
6	化合物别名	CGP-48933：Diovan
7	分子式	$C_{24}H_{29}N_5O_3$
8	分子量	435.53
9	化学结构（mol 文件）	

3　数据质量控制和评估

化学药物数据集为保证数据质量，在采集数据时制定了数据采集规范（见本文 1.2）。同时采用抽检的方式，随机抽选数据记录进行人工校对。为了解决数据的可追溯性问题，化学药物数据集在原始数据采集的同时录入数据来源号，来源号由 5 位数字编号，前 2 位数字为来源书籍的章号，第 3 位数字为节号，后 2 位数字为数据条目编号。由于工具书籍的编排具有严格的顺序性，因此可针对

数据集的连续性进行校验。在后续的数据处理中发现的数据遗漏或者数据质量问题，根据数据来源号对照原始数据得到了修正。

对数据集中的关键数据项，进行了程序校验。根据美国化学文摘社（CAS）发布的 CA 登记号有效性验证规范[9]，一个 CASRN 最多有 10 位数字，由连字符“-”分为三个部分，从左边起的第一部分的数字为 2 到 7 位数，第二部分数字为 2 位数，最后一部分由 1 位数组成。最后的一位数是校验码，数据集采用程序软件使用一个标准计算方法来计算 CAS 登记号是否为一个有效号码。

数据集中的化合物分子式和分子量数据，通过了分子式处理技术验证其精确性。如图 2 所示，程序读入分子式字符串 FormulaString 后解析其字符特征，比如分子式中的“（）”“.”等，以此判断该分子式是否为规范的表达，并且将分析所得的结果存入分子式特征变量 FormulaFeature 中；根据该分子式的类型，截取其中的分子式单元字符串，分别存入分子式单元字符串变量 FormulaUnit 中，普通的化合物分子式单元字符串为一个，晶体和有机盐分子式单元字符串为两个或多个，聚合物分子式单元为括号中单体分子式的字符串；依次遍历分子式单元字符串 FormulaUnit，分别解析出元素符号和元素数量存入元素变量 Atom 和元素数量变量 Amount 中，并且根据元素的原子量和元素数量计算该分子式基本单元的分子量，并且累加存入化合物分子量变量 MolecularWeight；最后输出化合物分子式特征数据 FormulaFeature 和化合物分子量数据 MolecularWeight。根据分子式数据处理程序所得的结果来验证数据集中的分子式是否符合规范，验证分子量数据是否正确。

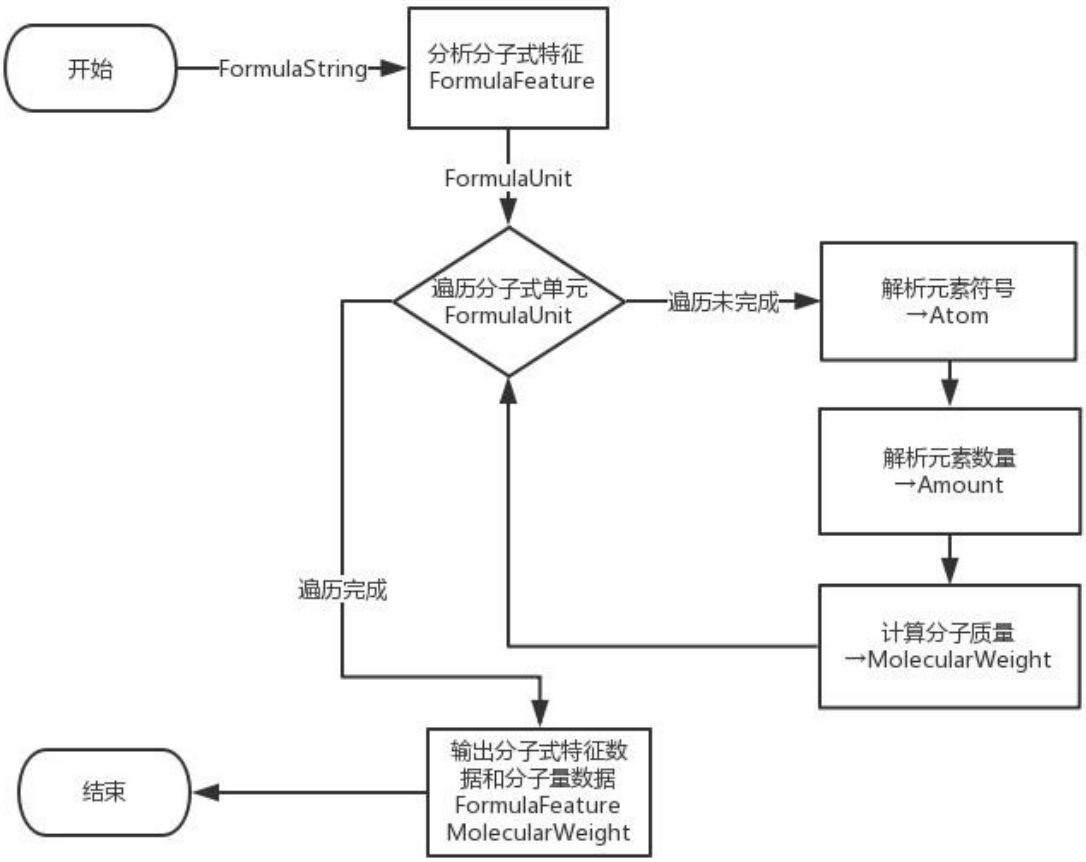


图 2　分子式数据处理程序算法

4 数据使用方法和建议

化学药物数据集使用 MDB 格式存储，使用者可以安装 Microsoft Office Access®软件，打开数据集并且对数据集进行查看、检索、数据挖掘等操作[10]。化学药物数据集中的数据描述具有一定的专业性，为了保证数据格式的统一性，数据中采用某些符号和缩写来代替相应的专业名词，符号和缩写说明如表 6 所示。

表 6　化学药物数据集符号和缩写说明表

序号	缩写符号	说明
1	$[\alpha]_D^{20}$	旋光度，下标为光源，上标为温度
2	n_D^{20}	折射率，下标为光源，下标为温度
3	d_4^{23}	相对密度，指一定体积的该物质在 23℃下的质量与等体积纯水在 4℃下的质量之比
4	$E_{1\ cm}^{1\%}$	消光度，下标为槽的厚度，上标为样品的浓度
5	ε	分子消光系数
6	mp	熔点
7	bp	沸点
8	pH	酸碱度
9	pKa	电解质电离常数倒数的对数
10	Ac	乙酰基
11	Bz	苯甲酰基
12	Bzl	苄基
13	Bu	丁基
14	Et	乙基
15	Me	甲基
16	Ph	苯基
17	Pr	丙基
18	Ts	对甲苯黄酰基
19	7-ACA	7-氨基头孢烷酸
20	7-ACT	7-氨基头孢三嗪
21	6-APA	6-氨基青霉烷酸
22	DCC, DCCD	二环己基二亚胺
23	DMA	N，N-二甲基乙酰胺
24	DMF	N，N-二甲基甲酰胺
25	DMSO	二甲基亚砜
26	THF	四氢呋喃
27	IR	红外光谱
28	UV	紫外光谱

序号	缩写符号	说明
29	USP	美国药典
30	DE	德国专利
31	EP	欧洲专利
32	GB	英国专利
33	JP	日本专利
34	US	美国专利
35	WO	世界专利

相对于工具书或者书籍只能根据目录和中英文索引来游览和检索，化学药物数据集的使用更为便捷和灵活。除了可以根据药物名称、性状、用途、制法等信息对药物进行检索和归类外，还可以利用数据集中药物化合物的化学结构信息进行量化计算。化学药物数据集包含的药物都有较为明确的化学结构，对于研究药物靶点信息、药物作用基团的构效关系等具有较好的数据支持作用，从而从数据角度为创新药物研发提供帮助。数据集中的信息描述言简意赅，收集的药物大都是已经市场化或广为人知的经典产品，适用于科研教学和大众科普教育等领域。化学药物数据集所收集药物的数据范围和数据量有限，但其建设方法和步骤具有一定的广谱性，后续可以本文所述加工处理方法为例，扩展到同领域其他数据源的加工处理，如《中国药典》《新编药物学》，补充增加化学药物数据集最新的数据源，进一步扩大数据覆盖范围。

数据作者分工职责

徐挺军（1984—），男，浙江海宁人，硕士，工程师，研究方向为化学数据库。主要承担工作：数据库设计和数据库建库。

赵英莉（1970—），女，辽宁沈阳人，硕士，副研究馆员，研究方向为化学信息学。主要承担工作：数据采集、基础数据加工和数据管理。

李英勇（1978—），男，河南南阳人，硕士，高级工程师，研究方向为化学信息学。主要承担工作：化合物数据登录。

参考文献

[1] 李广乾. 促进我国化学制药行业技术创新的政策研究[J]. 现代产业经济, 2013 (z1): 48-56.

[2] 孙大柠. 谈当今我国合成药物的创新研制与开发——访中国医学科学院药物研究所郭宗儒研究员[J]. 药学进展, 2010, 34(1): 1-6.

[3] 王可鉴, 石乐明, 贺林, 等. 中国药物研发的新机遇:基于医药大数据的系统性药物重定位[J]. 科学通报, 2014, 59(18): 1790-1796.

[4] 药品数据库[EB/OL]. http://www.organchem.csdb.cn/scdb/main/cdntd_introduce.asp.

[5] DrugBank version 5.1.1[EB/OL]. https://www.drugbank.ca/.

[6] 周学良. 精细化工产品手册.药物[M]. 北京: 化学工业出版社精细化工出版中心, 2003.

[7] 陈维明, 朱翠娣, 赵英莉, 等. 论数据唯一标识与化学数据的集成[C]. 第九届科学数据库与信息技术学术研讨会, 广西桂林, 2008.

[8] 赵英莉, 徐衍波, 李英勇, 等. 化合物参考数据库的设计[C]. 第十届科学数据库与信息技术学术研讨会, 贵州贵阳, 2010.

[9] American Chemical Society. Check Digit Verification of CAS Registry Numbers[EB/OL]. [2018-10-08]. http://www.cas.org/content/chemical-substances/checkdig.

[10] 纪澍琴, 李连德, 常耀辉. Access 数据库应用基础教程[M]. 北京: 北京邮电大学出版社, 2013.

论文引用格式

徐挺军, 赵英莉, 李英勇. 化学药物数据集[J/OL]. 中国科学数据, 2019, 4(1). (2018-11-22). DOI: 10.11922/csdata.2018.0061.zh.

数据引用格式

徐挺军, 赵英莉, 李英勇. 化学药物数据集[DB/OL]. Science Data Bank, 2018. (2018-10-08). DOI: 10.11922/sciencedb.657.

A dataset of chemical drugs

Xu Tingjun[1*], Zhao Yingli[1], Li Yingyong[1]

1. Information Center, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, P.R. China

*Email: xutingjun@sioc.ac.cn

Abstract: This study collected chemical drug data from book manuals. The collected data were then classified, summarized and standardized. Structured data of the drug compounds were integrated by using the method of compound unique identification. For quality control, we developed collection specifications and methods for data sampling and backtracking, which, coupled with algorithm programs, ensured the accuracy of the key data items. The dataset contains 1060 records that fall into two subsets: one for basic drug data and the other for drug compound data. This dataset provides data support for drug development, drug improvement, as well as relevant research and education, etc.

Keywords: drug data; chemical drugs; drug compounds; drug development

Dataset Profile

Title	A dataset of chemical drugs
Data corresponding author	Xu Tingjun(xutingjun@sioc.ac.cn)
Data authors	Xu Tingjun, Zhao Yingli, Li Yingyong
Time range	1985–2001
Geographical scope	Worldwide
Data volume	5.44 MB
Data format	*.MDB

Data service system	http://www.sciencedb.cn/dataSet/handle/657
Sources of funding	CAS informatization project during the Thirteenth Five-Year Plan (XXH135); Shanghai Chemistry & Chemical Industry Data Platform(18DZ2294000)
Dataset composition	This dataset consists of two parts of data, one for basic drugs (including their type, name, properties, preparation, application, manufacturer, etc.) and the other for drug compounds (including their registration number, CA registration number, name, molecular formula, molecular weight, chemical structure, etc.).