

Chuẩn hóa dữ liệu thuốc nhập khẩu

1. Kết quả thử nghiệm

Sau khi hoàn thiện các bước xử lý, kết quả thử nghiệm sơ bộ ban đầu AI tự điền được khoảng 76% dữ liệu trống. Tuy nhiên, các dữ liệu này là dữ liệu thô. Để có hiệu quả trong tương lai, cần tiếp tục xử lý để nâng cao chất lượng dữ liệu (sẽ trình bày ở mục 3).

Đầu vào : File dữ liệu thuốc trong nước.

STT	Loại giá	Ngày	Vào	Tên thuốc	Tên HC	NĐ/HL	Số GPLH/GPNK	Dạng bào chế	Quy cách đóng gói	ĐVT
	KK nhập khẩu	2020-10-03		Potassium Chloride	Kali chloride 1g/10ml	1g/10ml	VN-16303-13	Dung dịch đậm đặc	Hộp 50 ống PP	Ống
	KK nhập khẩu	2020-10-03		Nicardipine Ague	Nicardipin hydrochlor	10mg/10ml	VN-19999-16	Dung dịch tiêm	Hộp 10 ống x 10ml	Ống
	KK nhập khẩu	2020-10-03		Fluoride	Ofloxacin 3mg/1g		VN-16411-13	Mỡ tra mắt	Hộp 1 tuýp 3,5g	Tuýp
	KK nhập khẩu	2020-10-03		Bupivacaine Ague	Bupivacaine hydrochlor	100mg/20ml	VN-19692-16	Dung dịch tiêm	Hộp 10 lọ x 20ml	Lọ
	KK nhập khẩu	2020-10-03		NOZAXEN	Esomeprazol	40mg	VN-19598-16	Viên nén bao tan	Hộp 1 vỉ x 14 viên	Viên
	KK nhập khẩu	2020-10-03		Phenylephrine Ag	Mỗi 1ml dung dịch chứa	50 mcg	VN-21311-18	Dung dịch tiêm	Hộp 10 bơm tiêm đóng	Bơm tiêm
	KK nhập khẩu	2020-10-03		Jardiance	Empagliflozin 25mg	25mg	VN2-606-17	Viên nén bao phim	Hộp 3 vỉ x 10 viên	Viên
	KK nhập khẩu	2020-10-03		Trajenta	Linagliptin 5mg	5mg	VN-17273-13	Viên nén bao phim	Hộp 3 vỉ x 10 viên	Viên
	KK nhập khẩu	2020-10-03		Jardiance Duo	Mỗi viên chứa: Empag	12,5mg, 1000	VN3-185-19	Viên nén bao phim	Hộp 3 vỉ x 10 viên	Viên
	KK nhập khẩu	2020-10-03		Jardiance Duo	Mỗi viên chứa: Empag	12,5mg, 850m	VN3-186-19	Viên nén bao phim	Hộp 3 vỉ x 10 viên	Viên
	KK nhập khẩu	2020-10-03		Jardiance Duo	Mỗi viên chứa: Empag	5mg, 850mg	VN3-187-19	Viên nén bao phim	Hộp 3 vỉ x 10 viên	Viên
	KK nhập khẩu	2020-10-03		Jardiance Duo	Mỗi viên chứa: Metfor	1000mg, 5mg	VN3-188-19	Viên nén bao phim	Hộp 3 vỉ x 10 viên	Viên
	KK nhập khẩu	2017-09-21		Ursobil	Acid ursodeoxycholic	300mg	VN-20260-17		Hộp 3 vỉ x 10 Viên; Hộp	Viên
	KK nhập khẩu	2017-08-18		Neutromax 300m	Filgrastim	300mcg/ml	QLSP-0804-14		Hộp 1 lọ	Lọ
	KK nhập khẩu	2018-04-11		Voxin	Vancomycin (dưới dạng	1g	VN-20983-18		Hộp 1 lọ	Lọ
	KK nhập khẩu	2018-04-11		Liprillex	Lisinopril (dưới dạng Li	5mg	VN-20982-18		Hộp 3 vỉ x 20 Viên	Viên
	KK nhập khẩu	2018-05-08		Imazan	Azathioprine 50mg	50mg	VN-20726-17		Hộp 4 vỉ x 25 Viên	Viên
	KK nhập khẩu	2017-12-24		Colistimethate for	Colistin (dưới dạng Co	150mg	VN-20727-17		Hộp 1 lọ	Lọ
	KK nhập khẩu	2018-04-11		Acido Tranexamico	Acid tranexamic 500m	500mg	VN-20980-18		Hộp 5 ống 5ml	Ống
	KK nhập khẩu	2020-04-06		Tolsus	Mỗi 5ml chứa: Sulfam	200mg, 40mg	VN-22089-19	Hỗn dịch uống	Hộp 1 lọ 60ml	Lọ
	KK nhập khẩu	2020-10-03		Trajenta Duo	Linagliptin 2,5 mg; Me	2,5mg + 500m	VN3-5-16	Viên nén bao phim	Hộp 3 vỉ x 10 viên bao p	Viên
	KK nhập khẩu	2020-10-03		Trajenta Duo	Linagliptin 2,5 mg; M	2,5MG + 1000	VN3-4-16	Viên nén bao phim	Hộp 3 vỉ x 10 viên bao p	Viên
	KK nhập khẩu	2020-10-03		Redpharkit	Rabeprazole Sodium;	20mg/500mg/	VN-14839-12	Viên nén bao phim	Hộp 1 vỉ x 6 viên (mỗi l	Liều
	KK nhập khẩu	2020-10-03		Fegem-100	Sắt III hydroxyd dạng		VN-14829-12	Viên nén nhai	Hộp 10 vỉ x 10 viên	viên
	KK nhập khẩu	2020-09-30		ERANFU	Fulvestrant	250mg	VN3-259-20	Dung dịch tiêm	Hộp 2 bơm tiêm đóng s	Bơm tiêm
	KK nhập khẩu	2020-09-28		HemoQ Mom	Polysaccharide Iron c	326,1mg, 25m	VN-20490-17	Viên nang cứng	Hộp 3 vỉ x 10 viên	Viên
	KK nhập khẩu	2020-09-28		Hepagold	Mỗi 250ml chứa: L-Isol	2,25g, 2,75g, 4	VN-21298-18	Dung dịch tiêm	Hộp 10 vỉ x 10ml	Túi

Đầu ra : File dữ liệu được điền các thông tin cần thiết. (yêu cầu đạt được: điền được 70% số bản ghi trống)

KK nhập khẩu	2020-10-05	Jardiance	Empagliflozin 25mg	25mg	VN2-606-17	Viên nén bao phim	Hộp 3 vỉ
KK nhập khẩu	2020-10-05	Trajenta	Linagliptin 5mg	5mg	VN-17273-13	Viên nén bao phim	Hộp 3 vỉ
KK nhập khẩu	2020-10-05	Jardiance Duo	Mỗi viên chứa: Empag	12,5mg, 1000	VN3-185-19	Viên nén bao phim	Hộp 3 vỉ
KK nhập khẩu	2020-10-05	Jardiance Duo	Mỗi viên chứa: Empag	12,5mg, 850m	VN3-186-19	Viên nén bao phim	Hộp 3 vỉ
KK nhập khẩu	2020-10-05	Jardiance Duo	Mỗi viên chứa: Empag	5mg, 850mg	VN3-187-19	Viên nén bao phim	Hộp 3 vỉ
KK nhập khẩu	2020-10-05	Jardiance Duo	Mỗi viên chứa: Metfor	1000mg, 5mg	VN3-188-19	Viên nén bao phim	Hộp 3 vỉ
KK nhập khẩu	2017-09-21	Ursobil	Acid ursodeoxycholic	@300 mg	VN-20260-17	Viên nén	Hộp 3 vỉ
KK nhập khẩu	2017-08-18	Neutromax 300mc	Filgrastim	300mcg/ml	QLSP-0804-14	Dung dịch tiêm	Hộp 1 lọ
KK nhập khẩu	2018-04-11	Voxin	Vancomycin (dưới dạng	1g	VN-20983-18	Bột đông khô để pha o	Hộp 1 lọ
KK nhập khẩu	2018-04-11	Liprillex	Lisinopril (dưới dạng Li	5mg	VN-20982-18	Viên nén	Hộp 3 vỉ
KK nhập khẩu	2018-05-08	Imazan	Azathioprine 50mg	50mg	VN-20726-17	Viên nén bao phim	Hộp 4 vỉ
KK nhập khẩu	2017-12-26	Colistimethate for	Colistin (dưới dạng Co	150mg	VN-20727-17	Bột đông khô pha tiêm	Hộp 1 lọ
KK nhập khẩu	2018-04-11	Acido Tranexamico	Acid tranexamic 500m	500mg	VN-20980-18	Dung dịch tiêm truyền	Hộp 5 ố
KK nhập khẩu	2020-04-06	Tolsus	Mỗi 5ml chứa: Sulfam	200mg, 40mg	VN-22089-19	Hỗn dịch uống	Hộp 1 lọ
KK nhập khẩu	2020-10-03	Trajenta Duo	Linagliptin 2,5 mg; Me	2,5mg + 500m	VN3-5-16	Viên nén bao phim	Hộp 3 vỉ
KK nhập khẩu	2020-10-03	Trajenta Duo	Linagliptin 2,5 mg; M	2,5MG + 1000	VN3-4-16	Viên nén bao phim	Hộp 3 vỉ
KK nhập khẩu	2020-10-03	Redpharkit	Rabeprazole Sodium;	20mg/500mg/	VN-14839-12	Viên nén bao phim	Hộp 1 vỉ
KK nhập khẩu	2020-09-30	Fegem-100	Sắt III hydroxyd dạng	@	VN-14829-12	Viên nén nhai	Hộp 10
KK nhập khẩu	2020-09-30	ERANFU	Fulvestrant	250mg	VN3-259-20	Dung dịch tiêm	Hộp 2 b
KK nhập khẩu	2020-09-28	HemoQ Mom	Polysaccharide Iron c	326,1mg, 25m	VN-20490-17	Viên nang cứng	Hộp 3 vỉ
KK nhập khẩu	2020-09-28	Hepagold	Mỗi 250ml chứa: L-Isol	2,25g, 2,75g, 4	VN-21298-18	Dung dịch tiêm truyền	Thùng c
KK nhập khẩu	2020-09-28	Hepagold	Mỗi 250ml chứa: L-Isol	2,25g, 2,75g, 4	VN-21298-18	Dung dịch tiêm truyền	Thùng c
KK nhập khẩu	2020-09-28	Nephagold	Mỗi 250 ml chứa: L-Isol	1,4g, 2,2g, 1,6	VN-21299-18	Dung dịch tiêm truyền	Thùng c

Cụ thể đã điền được:

Dạng bào chế : 16.763 / 21.875 (76%)

Nồng độ hàm lượng : 4.443/5.782 (76%)

Tên hoạt chất : 633 / 988 (64 %)

2. Xây dựng cơ sở dữ liệu lớn phục vụ AI tham chiếu.


Thiết lập cơ sở dữ liệu lớn (bigdata) nhằm mục đích làm từ điển và dữ liệu nguồn sử dụng AI để tìm ra các thông tin thiếu như: tên hoạt chất, nồng độ hàm lượng, quy cách đóng gói.....

2.1 Nguồn từ các nước Châu âu:

Nguồn dữ liệu các loại lấy từ EMA (EUROPEAN MEDICINES AGENCY) , dữ liệu được phát hành vào tháng 7 năm 2018 được lấy từ 57 cơ sở dữ liệu của các khu vực thuộc EEA.

địa chỉ : <https://www.ema.europa.eu/en/>

Số lượng dữ liệu: 150.000 dữ liệu về tên thuốc, hoạt chất, nồng độ hàm lượng, quy cách đóng gói....

 EUROPEAN MEDICINES AGENCY SCIENCE MEDICINES HEALTH				
The information provided is based on data held in the Article 57 database. This data is submitted by holders of marketing authorisations (MAHs) for medicines in the European Union and the European Economic Area. MAHs must submit information to the European Medicines Agency (EMA) on authorised medicines and keep this information up-to-date in accordance with EU pharmaceutical regulation. The aim of the publication is to disseminate information on the locations where pharmacovigilance system master files are kept and the contact information for pharmacovigilance enquiries. Further information on the Article 57 database can be found here: https://www.ema.europa.eu/en/human-regulatory/post-authorisation/data-medicines-iso-idmp-standards/data-submission-authorised .				
Product name ²	Active substance	Route of administration	Product authorisation country	Marketing authorisation
Product short name: brand name or the combination of the generic name and the company name	The symbol "I" is used to separate different pharmaceutical products			
A 313	Vitamin A Concentrate (Oily Form)	Cutaneous Use	France	Laboratoires Pharma Developper
A 313	Vitamin A Concentrate (Oily Form)	Oral Use	France	Laboratoires Pharma Developper
A L E R I D	Cetirizine Dihydrochloride	Oral Use	Czech Republic	Cipla Europe Nv
A R G O T O N E	Ephedrine Hydrochloride, Silver Pro	Nasal Use	Italy	Dompé Farmaceutici S.P.A.
A.T. 10 Perlen	Dihydrotachysterol	Oral Use	Germany	Teofarma S.R.L.
A.Vogel Cystoforce Blaasformule	Tincture From Purple Coneflower HP	Oral Use	Netherlands	A.Vogel B.V.
A.Vogel Hyperiforce	Hypericum Perforatum L.	Oral Use	Netherlands	A.Vogel B.V.
Aaa Sore Throat	Benzocaine Ph. Eur.	Oromucosal Use	Malta	Manx Pharma Ltd

2.2 Nguồn từ Anh:

eMC : ELECTRONIC MEDICINES COMPENDIUM

là website thông tin về thuốc được thành lập từ 1999, cung cấp thông tin cập nhật, chính xác về các loại thuốc đang được lưu hành hợp pháp trên thị trường UK

địa chỉ : <https://www.medicines.org.uk/emc/browse-ingredients>

Số lượng dữ liệu: 2000 dữ liệu về thuốc, hoạt chất đang lưu hành tại Anh.

Browse active ingredients

Active ingredients beginning with A

0-9 **A** B C D E F G H I J K L M N O P Q R S T U V W X Y Z

> abacavir	> amiodarone hydrochloride
> abacavir hydrochloride	> amisulpride
> abacavir sulfate	> amitriptyline hydrochloride
> abatacept	> amlodipine besilate
> abciximab	> amlodipine maleate
> abemaciclib	> amlodipine mesilate monohydrate
> abietis oil	> ammonia liquor
> abiraterone acetate	> ammonia solution, aromatic
> acalabrutinib	> ammonia solution, strong
> acamprostate	> ammonium carbonate
> acamprostate calcium	> ammonium chloride
> acarbose	> ammonium salicylate
> acebutolol hydrochloride	> amorolfine hydrochloride
> aceclofenac	> amoxicillin
> acemetacin	> amoxicillin sodium

2.3 Danh sách từ điển các cách bào chế thuốc của nước ngoài (cụ thể : Mỹ)

Thông tin được lấy từ trang y tế chính thức của chính phủ Mỹ.

Nguồn : <https://www.fda.gov/>

Dosage Forms

[f Share](#) [t Tweet](#) [in LinkedIn](#) [✉ Email](#) [🖨 Print](#)

NCI Thesaurus OID: 2.16.840.1.113883.3.26.1.1

NCI concept code for pharmaceutical dosage form: C42636

SPL Acceptable Term	Code
AEROSOL	C42887
AEROSOL, FOAM	C42888
AEROSOL, METERED	C42960
AEROSOL, POWDER	C42971
AEROSOL, SPRAY	C42889
BAR, CHEWABLE	C42892
BEAD	C42890
CAPSULE	C25158
CAPSULE, COATED	C42895
CAPSULE, COATED PELLETS	C42896
CAPSULE, COATED, EXTENDED RELEASE	C42917
CAPSULE, DELAYED RELEASE	C42902
CAPSULE, DELAYED RELEASE PELLETS	C42904
CAPSULE, EXTENDED RELEASE	C42916
CAPSULE, FILM COATED, EXTENDED RELEASE	C42928

Số lượng dữ liệu: 150 dạng bào chế thuốc khác nhau.

2.4 Dữ liệu thuốc trong nước :

Nguồn : drugbank.vn

Số lượng dữ liệu: 12.000 bản ghi về hoạt chất, tên thuốc....

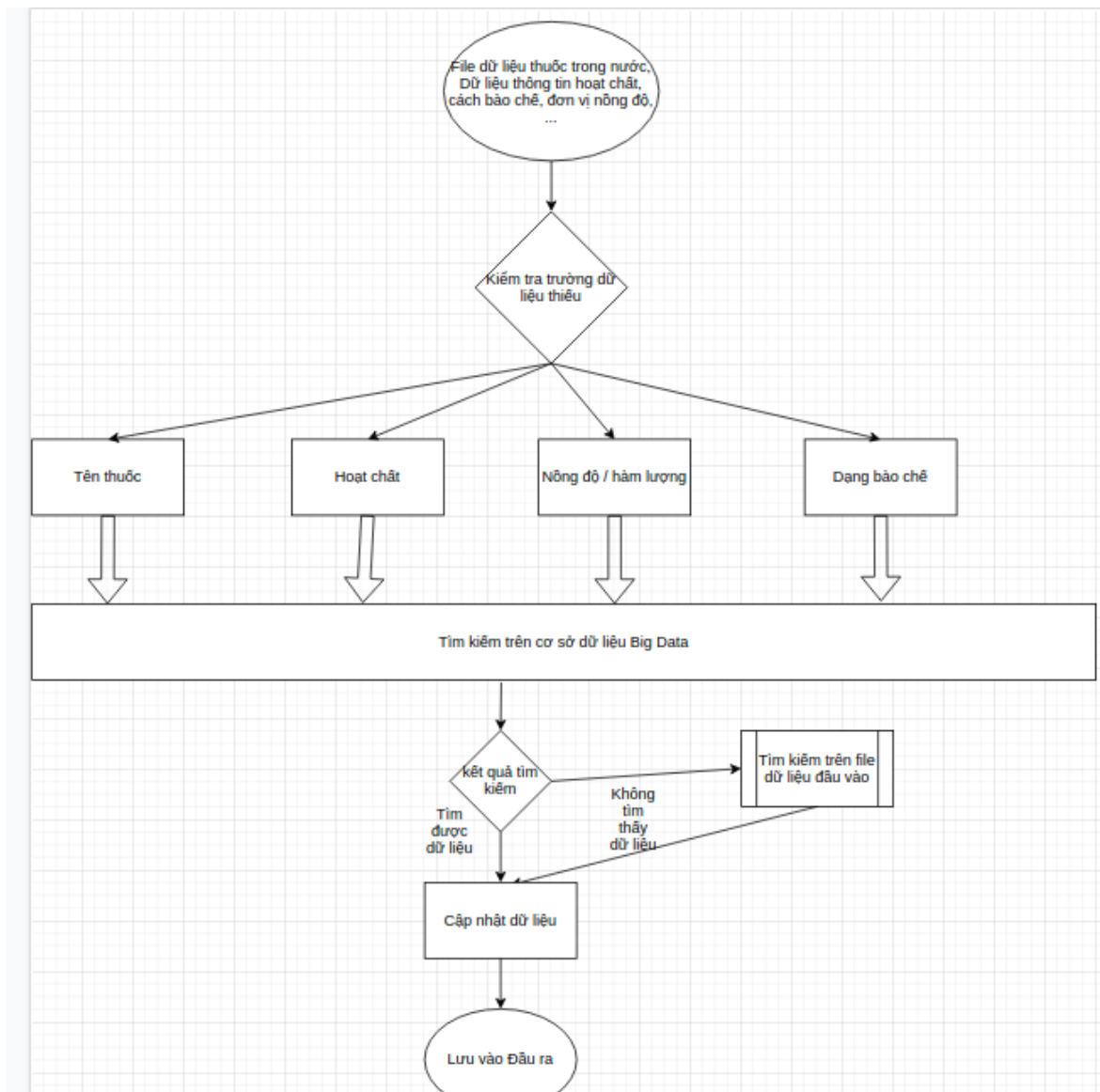
3. Training AI

Sử dụng dữ liệu bigdata đã thu thập để training AI. Mục đích để AI có thể phân tích và trợ giúp chuẩn hóa dữ liệu:

Các thông tin đưa cho AI đọc:

- + Tên thuốc
- + Tên hoạt chất
- + Nồng độ / hàm lượng
- + Dạng bào chế

- Cách thu thập dữ liệu còn thiếu:



3.1 Cách chuẩn hóa dữ liệu:

Đầu vào : Dữ liệu thuốc trong nước đã được điền các thông tin cần thiết

Đầu ra : Dữ liệu được chuẩn hóa và lưu lại vào cơ sở dữ liệu



3.2 Các vấn đề gặp phải trong file dữ liệu thuốc trong nước:

+ Sử dụng ngôn ngữ tự nhiên trong một số thuốc:

Ví dụ:

Tên thuốc: Gardasil

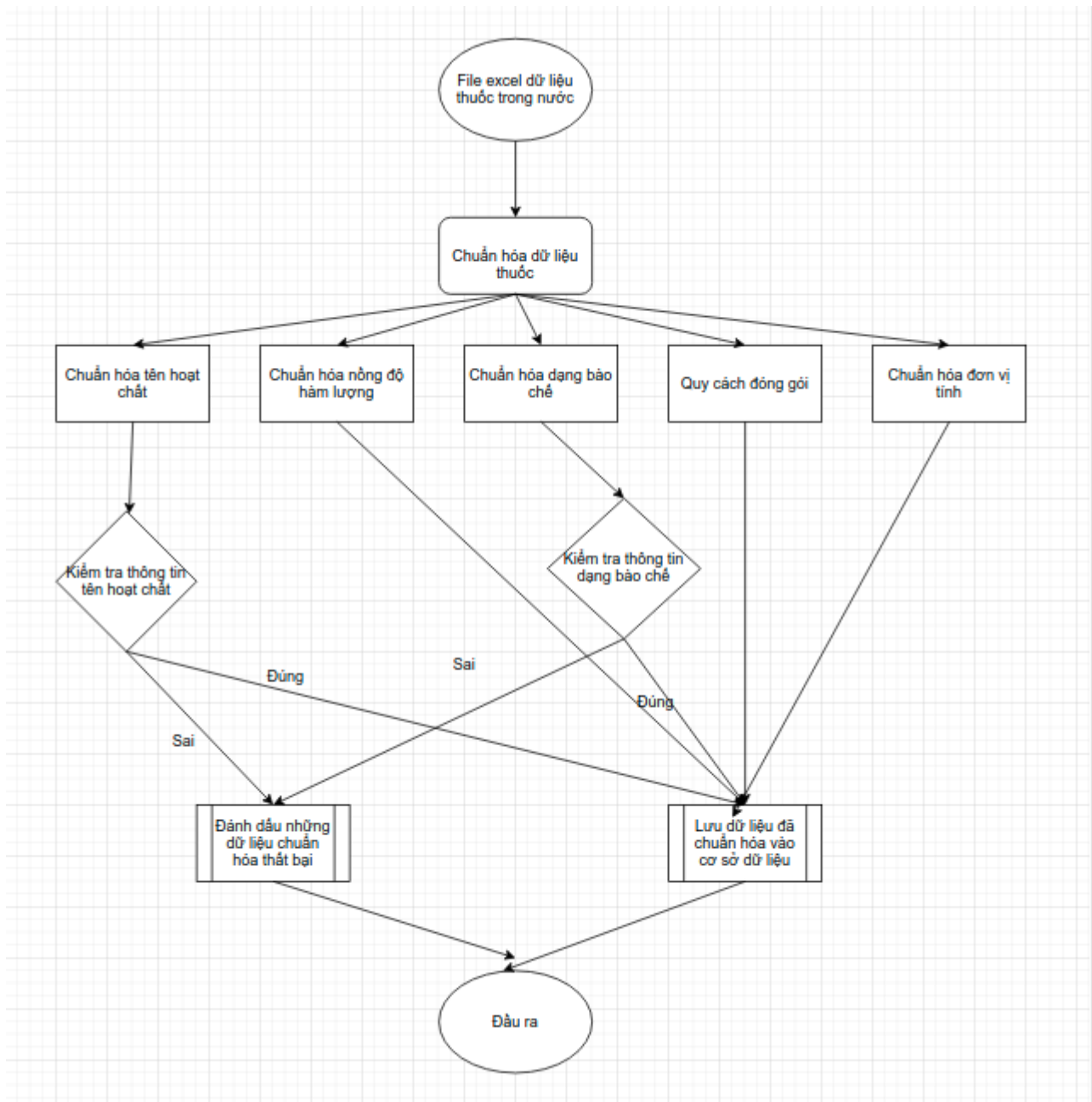
Tên hoạt chất : Vắc xin tái tổ hợp tứ giá phòng vi-rút HPV ở người tít 6,11,16,18. Mỗi liều 0,5 ml chứa 20 mcg protein L1 HPV6, 40 mcg protein L1 HPV11, 40 mcg protein L1 HPV16, 20 mcg protein L1 HPV18

Nồng độ / Hàm lượng: Mỗi liều 0,5 ml chứa 20 mcg protein L1 HPV6, 40 mcg protein L1 HPV11, 40 mcg protein L1 HPV16, 20 mcg protein L1 HPV18

...

+ Một số trường dữ liệu còn khuyết, không có dữ liệu: dạng bào chế, nồng độ hàm lượng, ...

3.3 Giải pháp:



3.3.1 Chuẩn hóa tên hoạt chất:

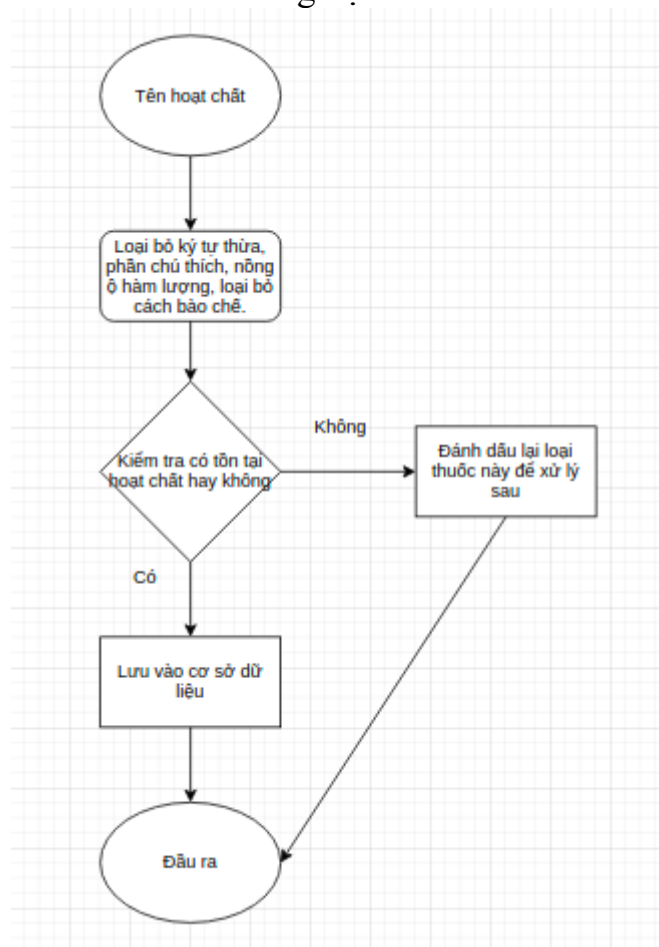
Trong dữ liệu thuốc Việt Nam, tên hoạt chất thường được viết kết hợp với nồng độ hàm lượng. Ngoài ra thường được chú thích kèm theo và sử dụng một số ngôn ngữ tự nhiên bổ sung.

Một số kiểu viết phổ biến như :

<tên hoạt chất><nồng độ><cách bào chế>

<tên hoạt chất><nồng độ>

<tên hoạt chất><cách bào chế><nồng độ>



Tại bước kiểm tra tên hoạt chất tồn tại : tìm kiếm tên hoạt chất trong cơ sở dữ liệu Big Data để tìm kiếm tên quốc tế của tên hoạt chất n

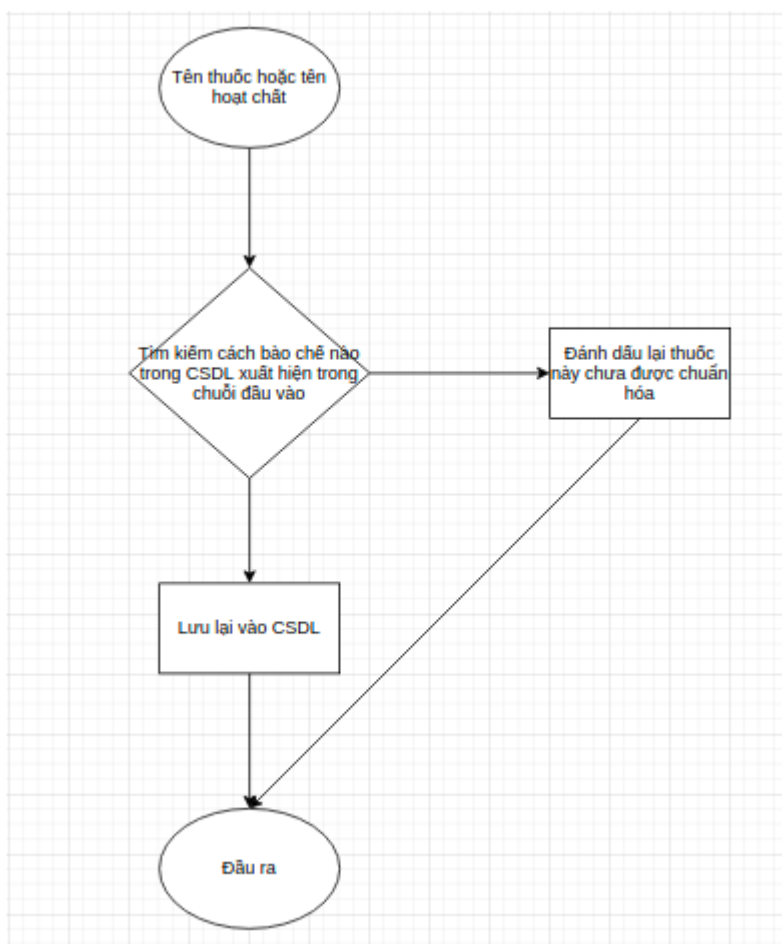
3.3.2 Chuẩn hóa hàm lượng nồng độ:

Trường hợp bản ghi không có nồng độ / hàm lượng cụ thể. Mà nồng độ hàm lượng lại nằm trong tên hoạt chất sẽ sử dụng cắt chuỗi dựa vào một số dấu hiệu các mẫu hàm lượng nồng độ điển hình như:

số + đơn vị ;
số + đơn vị / số + đơn vị
số + đơn vị / đơn vị

3.3.3 Chuẩn hóa dạng bào chế:

Trường hợp bản ghi không có dạng bào chế, dạng bào chế thường xuất hiện ở trong tên thuốc hoặc tên hoạt chất.



3.3.4 Chuẩn hóa quy cách đóng gói:

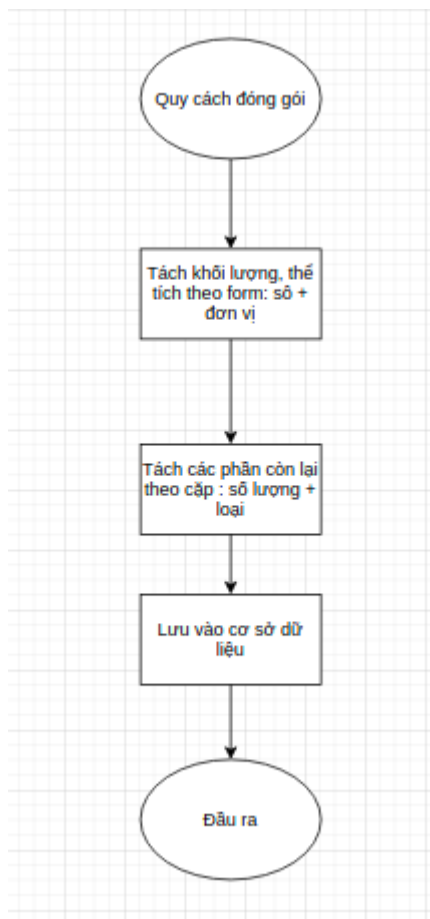
Cách đóng gói có nhiều cách khác nhau nhưng thường được liên kết với nhau bằng dấu “x”:

Ví dụ :

Hộp có 3 vỉ x 5 viên 500mg, Hộp 5 vỉ x 5 viên 700mg

Chuẩn hóa thành như sau :

[
 {“vỉ”: 3, “viên”: 5, “khối lượng”: “500mg”},
 {“vỉ”: 5, “viên”: 5, “khối lượng”: “700mg”}
]



3.3.5 Chuẩn hóa đơn vị tính:

Tìm kiếm đơn vị trong cơ sở dữ liệu Big Data. Thực hiện đối chiếu để tìm loại đơn vị quốc tế của nó.