Topic Medical Concept Embedding: Multi-Sense Representation Learning for Medical Concept

Feng Qian 1 *, Chengyue Gong * 3, Luchen Liu 1, Lei Sha2, Ming Zhang 1 †

1 Institute of Network Computing and Information Systems
2 Key Laboratory of Computational Linguistics, Ministry of Education School

EECS, Peking University

3 Department of Information Management

Peking University

* co-author and contributed equally† corresponding author

Terms

- [medical concept]: medical entities

- Concludes：
  - Medications：Such as Aspirin, penicillin
  - Medical signs：heart rate, body temperature

Terms

- [Symptom]：symptom, what leads to various [medical concepts].

- [Symptom] can be “fever”, “blood-related”, “alleviation” etc. general things.

Terms

- [distributed representation]

- Low dimensional vector representations for real work medical concepts.

Terms

- [concept-dosage pair]
- A same [medical concept] can pair with different dosages. Eg. 50 mg Aspirin or 100 mg Aspirin
- We call such pair [concept-dosage pair]

Motivations

- Learning [distributed representation] for [medical concept] can help us:
  - Better understanding about relationships behind [medical concepts]
  - Encode rich information in [distributed representation], and help other machine learning tasks and models in medical area.

Problem found

- Obviously, a lot of [medical concepts] are pointing to more than one [symptoms]. Eg.
  - Eg., Aspirin ([medical concept]) can cure both fever ([symptom]) and cardiovascular disease (another [symptom])
- So that multi-sense [distributed representation] learning is needed

Difference brought by dosage

- [medical concepts] are always accompanied by dosage information.
- Eg, [Aspirin-50mg], [80bpm-heart rate]

- Such dosage information can not be utilized in previous works.

Intuitions

- Traditional [distributed representation] learning models such as skip-gram learns information from data context. That is to say the co-occurrance between concepts in data.

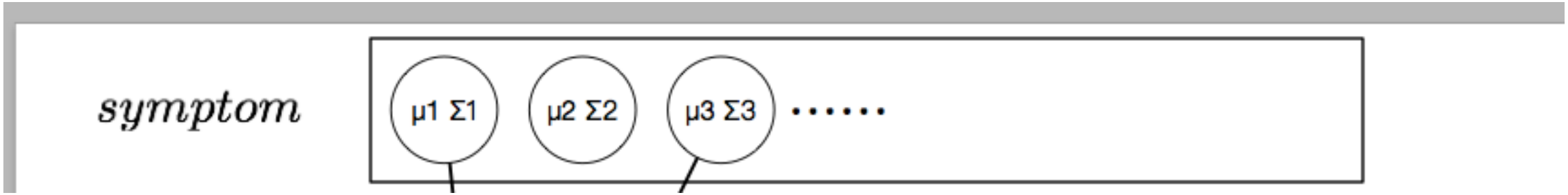
- Topic model can:
  - Assign different [symptoms] to a same [medical concept], so that it can help realize multi-sense representation learning.
  - In topic model, [symptoms] can be shared among [medical concepts]. This is also intuitively correct, sine a lot [concepts] points to a same [diagnose].

Topic Medical Concept Embedding (TMCE): Multi-Sense Representation Learning for Medical Concept

- TMCE combines both intuitions
- TMCE uses topic model to construct N-N relationships between [medical concepts] and [symptoms], at the same time, utilize information from context.
- We also considered dosage information into the inference process

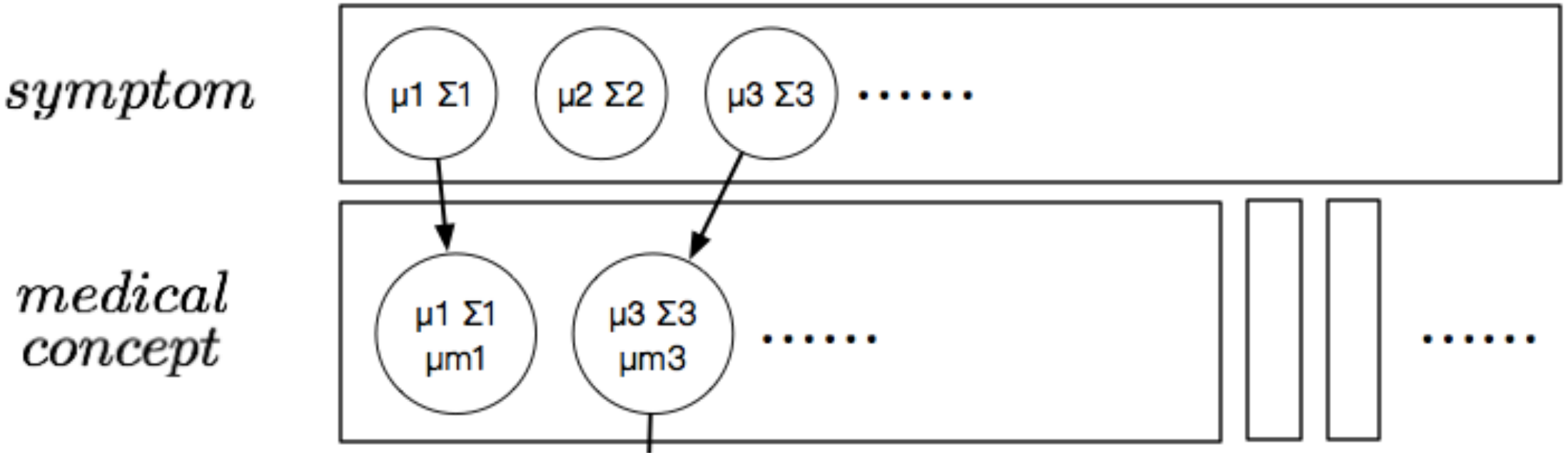
Model

- Firstly a Dirichlet process are designed to represent and generate all [symptoms]
- A circle in the picture stands for a [symptom], represented by a normal distribution



Model

- The top layer (symptom layer) DP is used as the base distribution for lower layer (medical layer).
- Each [medical concept] are sampled from top layer DP, and may contain more than one [symptom]. (multi-sense)
- So that [Symptoms] are shared by [medical concepts]



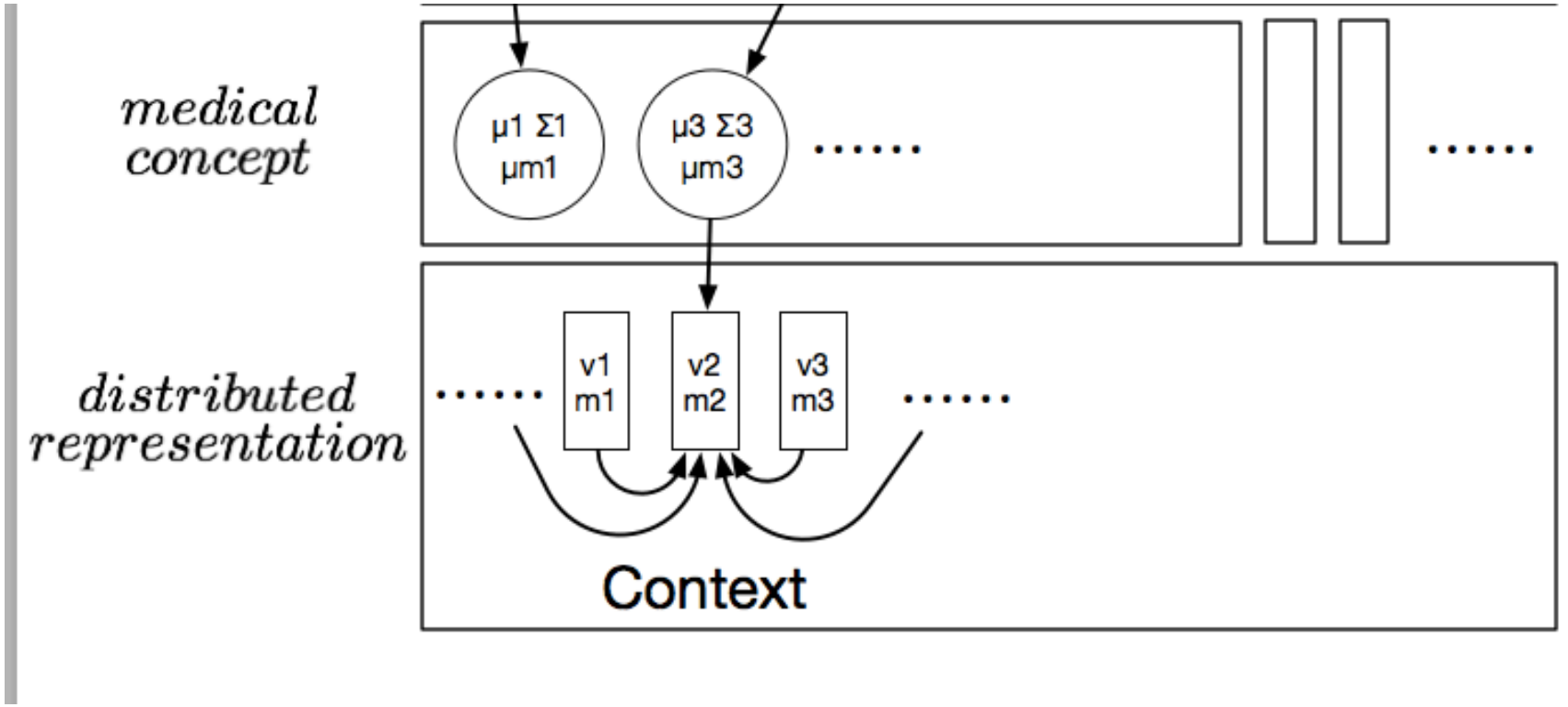
A further assumption

- When a specific [medical concept] is related to a specific [symptom], the dosage of the [medical concept] should follow normal distribution around a fixed dosage.

- Eg.， when Aspirin ([medical concept]) are used to cure fever([symptom]), the dosage should be around x mg with a variance of y, normal distribution.

模型

- Finally, [distributed representations] for [concept-dosage pairs] are generated under the influence of both topic model and context (skip gram)



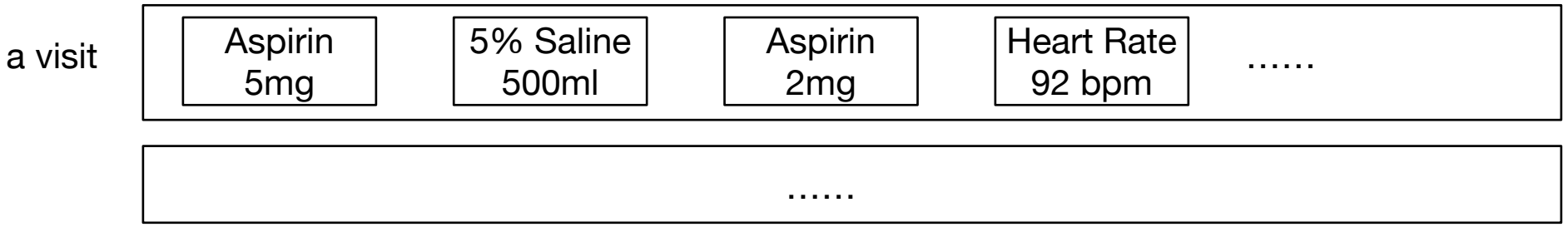
Practical Training

- For sake of high training efficiency, we divided the training process into two steps:

- 1- Train [distributed representation] with skip gram
- 2- Fix all [distributed representations], inference latent parameters of topic model。

- Iteration…

Quantitative experiment



- In each visit in dataset, there are [diagnose] assigned by doctors。If a [concept-dosage] appeared in this visit, [diagnoses] assigned to this visit are used as tags for this [concept-dosage pair]
- Binary classification task
- Predict [diagnose] by representation of [concept-dosage pair]

Result

		10%	20%	30%	40%	50%	60%	70%	80%	90%
macro_f	ours	**0.68628281**	**0.6956575**	**0.70963687**	**0.72387596**	**0.73831183**	**0.75510868**	**0.77031838**	**0.78519718**	**0.79866482**
	skip_gram_with_dose	0.67072476	0.6730499	0.67692377	0.68030946	0.68596564	0.6933448	0.70385601	0.71814867	0.73936982
	skip_gram_without_dose	0.67345436	0.67330197	0.68030007	0.69216314	0.69237111	0.69703536	0.71090192	0.71834783	0.71250866
	stacked_autoencoder	0.65672653	0.65139108	0.65496888	0.66955329	0.67448598	0.68156733	0.70399234	0.71346545	0.7374531
micro_f	ours	**0.686595**	**0.69623562**	**0.71075269**	**0.72574616**	**0.74129631**	**0.7593038**	**0.77537472**	**0.79120666**	**0.80523281**
	skip_gram_with_dose	0.67089141	0.67322947	0.67711731	0.68049233	0.68612577	0.69350899	0.70403852	0.71833481	0.73962466
	skip_gram_without_dose	0.67469053	0.67439933	0.68119604	0.69269972	0.69307127	0.69810097	0.71193247	0.71935878	0.7136105
	stacked_autoencoder	0.65674084	0.65166519	0.65508516	0.67015676	0.67501838	0.68187357	0.70428717	0.7139447	0.73819438

- X axis stands for proportion of training data.
- We used MIMIC III data by MIT.
- Distributed representation learned by TMCE can predict diagnose better than other baselines.

Case study-1

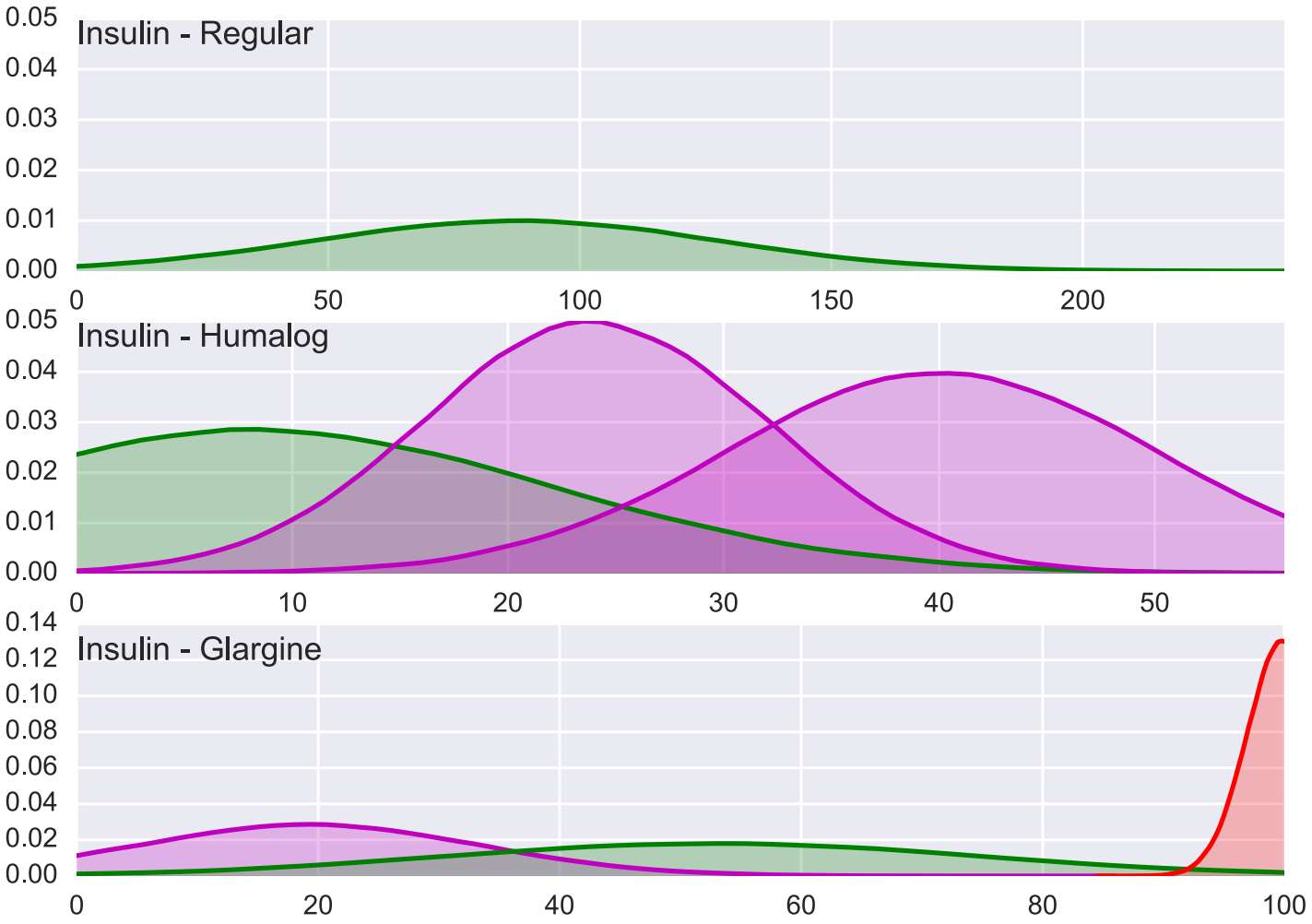
- [Medical concepts] under a same [diagnose] are highly related

Cluster 2:血红细胞
Intubation
PLATELET COUNT
PO2
BICARBONATE
Blood Cultured
RED BLOOD CELLS
BASE EXCESS
ANION GAP
PHOSPHATE
RDW
HEMOGLOBIN
HEMATOCRIT
UREA NITROGENRDW

Cluster 18:营养液
CALCIUM
LIPASE
Po Intake
Invasive Ventilation
BILIRUBIN
O2 FLOW
OR Crystalloid
PO Intake
FIBRINOGEN
Arterial Line
INR(PT)
D5 1/2NS
Calcium Gluconate
Dextrose 5%
TRIGLYCERIDES
Propofol
Packed Red Blood Cells
Heparin
OSMOLALITY

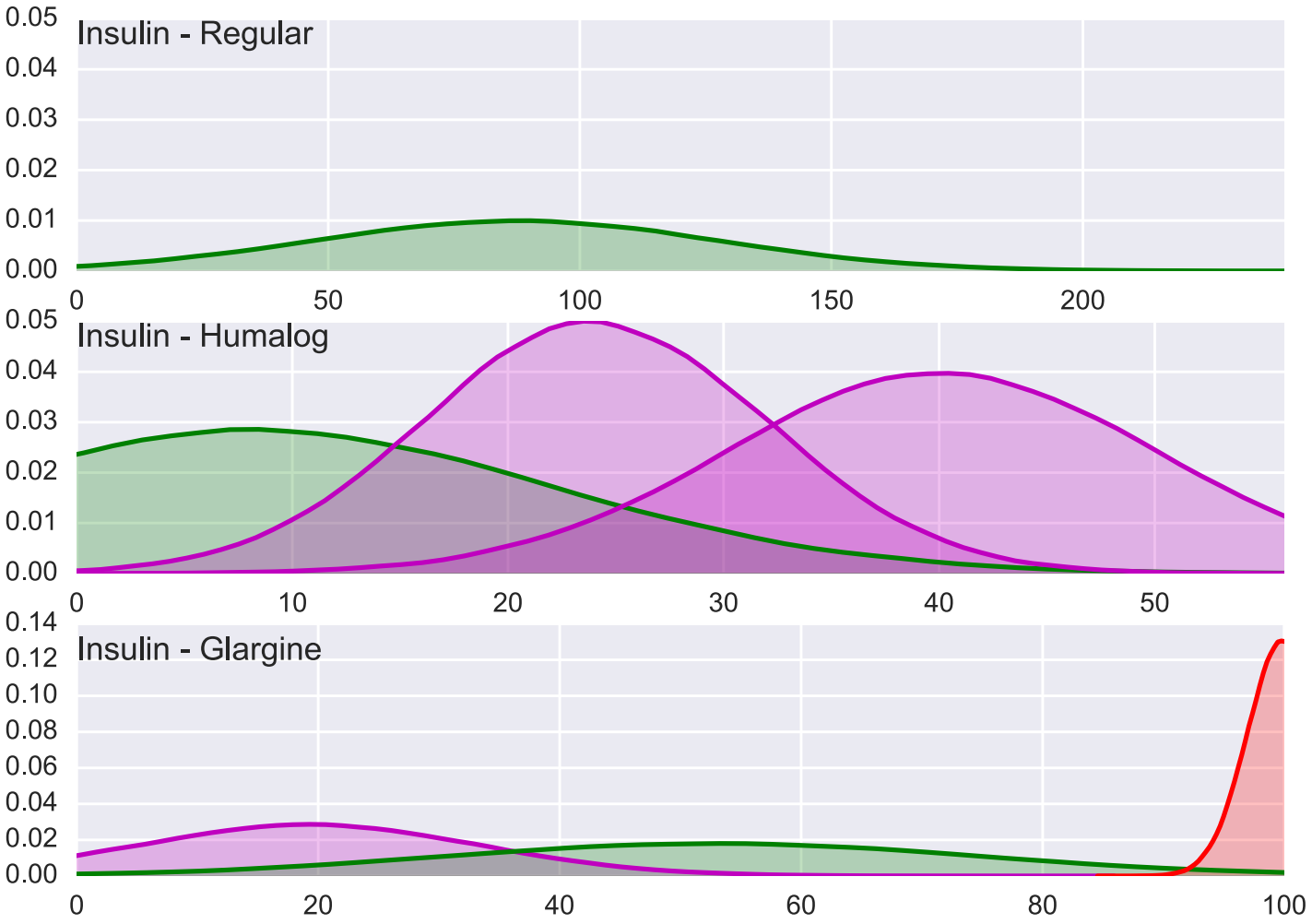
Cluster 18:胰岛&镇静剂
Insulin - Humalog
Insulin – Glargine
Dextrose 5%
Insulin – Regular
morphine sulfate
Lorazepam (Ativan)
neo-synephrine
Norepinephrine
Propofol

Case study-2, concept relationships

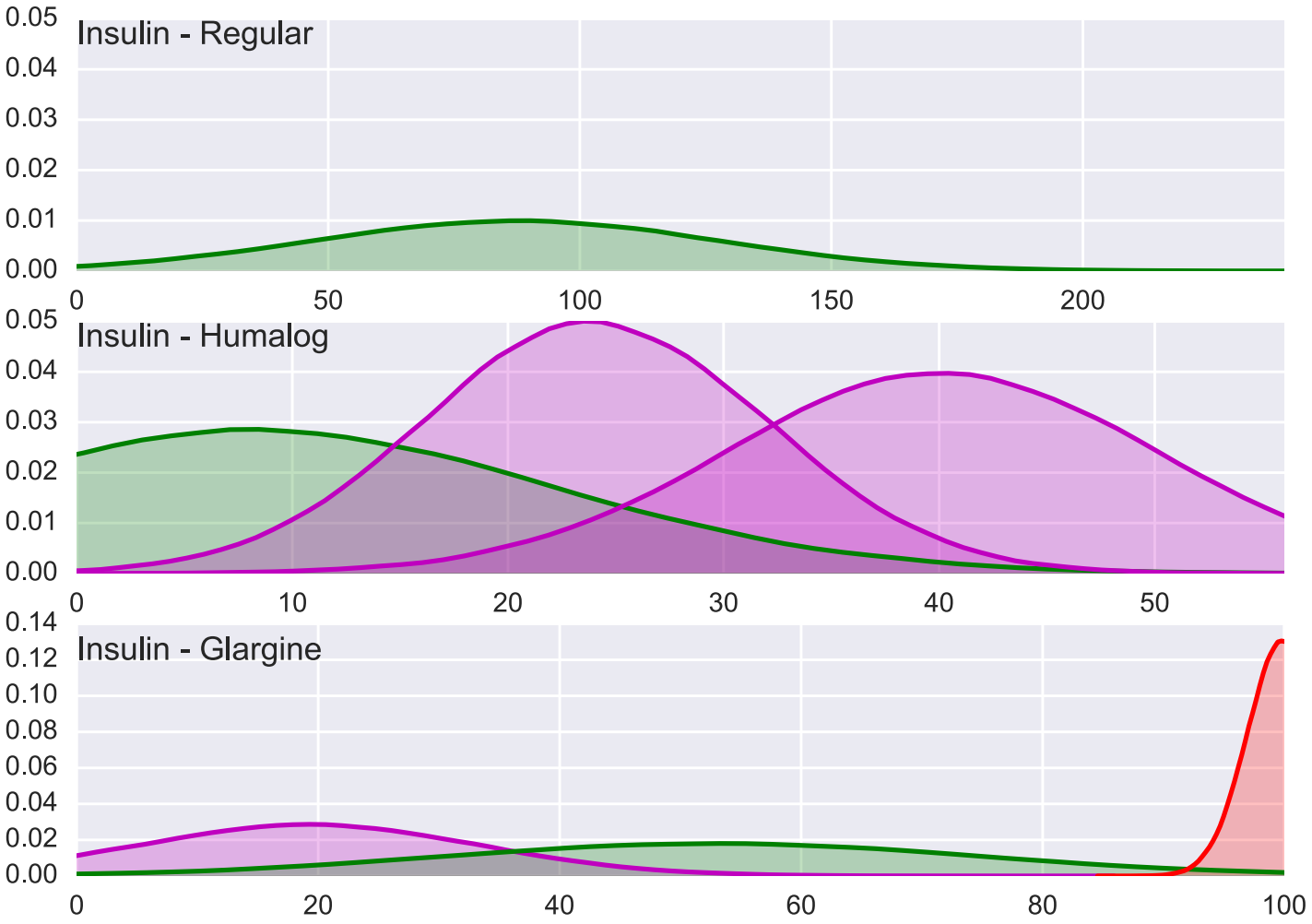


- Three plots stand for three kinds of diabetes drug
- The first one is a regular diabetes drug. The second one is a fast rate strong diabetes drug. The last one is a slow rate diabetes drug

Case study-2, concept relationships



- Same color stands for same global [symptom]
- All three kinds of diabetes drug are assigned similar [symptoms]



- If we look at the green global symptoms in the three plots, we will find:
- To gain a similar effect, fast rate drug needs lower dosage while slow rate drug needs higher dosage.
- This case shows that dosage information are successfully taken into account.

Thanks for your time and patience!