DEVELOPING AN EVIDENCE MATCHING FRAMEWORK USING WEB-BASED MEDICAL LITERATURE

JOYCE HO
COMPUTER SCIENCE DEPARTMENT
@joycehoUT
INTRODUCTION

EHR-BASED PHENOTYPING

- A collection of clinical characteristics present in EHRs that define a patient group or condition of interest

- Uses
  - Population understanding
  - Targeted screening & interventions
  - Patient identification for RCTs

INTRODUCTION

HIGH-THROUGHPUT EHR-BASED PHENOTYPING: RECENT DEVELOPMENTS

EHR database → Machine learning algorithms → Phenotypes

Mild Hypertension
(31.1% of patients)
- Hypertension
- ACE Inhibitors
- Thiazides and Thiazide-Like Diuretics

Phenotype 4
- Phenotype importance
- Diagnosis factor
- Medication factor
- Patient factor

Phenotype 1

Phenotypes
HIGH-THROUGHPUT EHR-BASED PHENOTYPING: RECENT DEVELOPMENTS

INTRODUCTION

HOW TO EVALUATE RESULTING PHENOTYPES?
MOTIVATION FOR EXTERNAL VALIDATION

Raw Electronic Healthcare Record Data

Automatic Phenotype Generation Process

Verification by Panel of Experts

Candidate Phenotypes

Verified Phenotypes

<table>
<thead>
<tr>
<th>Phenotype 2</th>
<th>Phenotype n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Symptom, Anomalities (v)</td>
<td>Iron Deficiency and Other/Unspecified Anemias and Blood Disease (v)</td>
</tr>
<tr>
<td>Heart Rhythm (v)</td>
<td>Other Endocrine Disease (v)</td>
</tr>
<tr>
<td>Other Respiratory (v)</td>
<td>Chronic Obstructive Pulmonary Disease (v)</td>
</tr>
<tr>
<td>Surgical Musculoskeletal (v)</td>
<td>Other Gastrointestinal Disorders (v)</td>
</tr>
<tr>
<td>Osteoarthritis of Hip or Knee (v)</td>
<td>Hypertension (v)</td>
</tr>
<tr>
<td>Surgical Procedures on the Musculoskeletal System (v)</td>
<td>Surgical Procedures on the Musculoskeletal System (v)</td>
</tr>
</tbody>
</table>
INTRODUCTION

MOTIVATION FOR EXTERNAL VALIDATION

Raw Electronic Healthcare Record Data

Automatic Phenotype Generation Process

Verification by Panel of Experts

DISAGREEMENTS & TIME-INTENSIVE

Candidate Phenotypes

Verified Phenotypes

- Major Symptoms, Abnormalities
- Heart Arrhythmias
- Other or Unspecified Anemias and Blood Disease
- Chronic Obstructive Pulmonary Disease
- Other Gastrointestinal Disorders
- Osteoarthritis of Hip or Knee
- Hypertension
- Surgical Procedures on the Musculoskeletal System

Phenotype 2

Phenotype n
“This isn’t my specialty – have you thought of looking at PubMed?”
SEARCHING WEB-BASED MEDICAL LITERATURE

“This isn’t my specialty – have you thought of looking at PubMed?”

“That’s a great suggestion! We should explore this direction.”
INTRODUCTION

SEARCHING WEB-BASED MEDICAL LITERATURE

“This isn’t my specialty – have you thought of looking at PubMed?”

“That’s a great suggestion! We should explore this direction.”

“…we extracted all articles from PubMed with the term “birth month” and an additional article referenced by a located article (n=156). We manually reviewed all abstracts and removed articles...This process identified 92 relevant articles.”—Boland et al. JAMIA 2015
“This isn’t my specialty – have you thought of looking at PubMed?”

“That’s a great suggestion! We should explore this direction.”

“…we extracted all articles from PubMed with the term “birth month” and an additional article referenced by a located article (n=156). We manually reviewed all abstracts and removed articles...This process identified 92 relevant articles.” —Boland et al. JAMIA 2015

“Although we might have expected the total time to be zero when there were no citations, we noted that there was a certain amount of start-up time even when there were few citations. The quadratic equation is as follows: $Total \ time = 721 + 0.243x - 0.0000123x^2$, where $x$ denotes the number of citations before exclusion criteria are applied. The predicted start-up time is 721 hours.”

—Allen et al. JAMA 1999
PHEKNOW-CLOUD/PIVET: A TOOL FOR VALIDATING PHENOTYPES VIA PUBMED

Phenotypes from different sources

High-Throughput Phenotypes

Peer-Reviewed Paper

Phenotype KnowledgeBase

PheKB

MongoDB

Phenotype Database

Phenotypic Item Representation

Corpus Analysis

Clinical Validity Determination

Phenotype Evidence Results


OVERALL PROCESS

Phenotype 33
- cardiac dysrhythmias
- heart failure
- atrial fibrillation and flutter
- unspecified chest pain
- cardiomyopathy
- calcium channel blocking agents
- loop diuretics
- antianginal agents

Phenotype

Generation of Candidates for "Synonym Set"

"heart failure", "heart failure", "congestive heart disease", "myocardial failure", ...

Ranking synonyms and related concepts

"antianginal agents", "amiodarone hydrochloride", "form amiodarone", "amiodarone 300mg 10ml solution"

Ranking synonyms and related concepts

Pubmed Articles

- "heart failure", "antianginal agents" -> 57
- "heart failure", "vitamins" -> 1.05

Ranked n-grams for Phenotypic items

Co-occurrence search in Pubmed

Lift for all co-occurrences
OVERALL PROCESS

PHENOTYPIC SYNONYM GENERATION

CO-OCCURRENCE ANALYSIS
**PHENOTYPIC SYNONYM GENERATION (MESH)**

**Candidate Synonyms**

<table>
<thead>
<tr>
<th>Tree Number</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>C08.381.423.847</td>
<td>Familial Primary Pulmonary Hypertension</td>
</tr>
<tr>
<td>C14.907.489.330</td>
<td>Hypertension, Malignant</td>
</tr>
<tr>
<td>C14.907.489.480</td>
<td>Hypertension, Pregnancy-Induced</td>
</tr>
<tr>
<td>C08.381.423</td>
<td>Hypertension, Pulmonary</td>
</tr>
<tr>
<td>C14.907.489.631</td>
<td>Hypertension, Renal</td>
</tr>
<tr>
<td>C14.907.489.631.485</td>
<td>Hypertension, Renovascular</td>
</tr>
<tr>
<td>C14.907.489</td>
<td>Hypertension</td>
</tr>
<tr>
<td>C14.907.489.861</td>
<td>Masked Hypertension</td>
</tr>
<tr>
<td>C13.703.395.249</td>
<td>Pre-eclampsia</td>
</tr>
<tr>
<td>C14.907.653</td>
<td>Prehypertension</td>
</tr>
<tr>
<td>C18.452.648.861.770</td>
<td>Pseudohypoaldosteronism</td>
</tr>
<tr>
<td>C14.907.489.907</td>
<td>White Coat Hypertension</td>
</tr>
</tbody>
</table>

**Ranked Synonyms**

<table>
<thead>
<tr>
<th>Name</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension</td>
<td>1.000</td>
</tr>
<tr>
<td>Hypertension, Malignant</td>
<td>0.600</td>
</tr>
<tr>
<td>White Coat Hypertension</td>
<td>0.400</td>
</tr>
<tr>
<td>Prehypertension</td>
<td>0.262</td>
</tr>
<tr>
<td>Hypertension, Renal</td>
<td>0.243</td>
</tr>
<tr>
<td>Pseudohypoaldosteronism</td>
<td>0.167</td>
</tr>
<tr>
<td>Masked Hypertension</td>
<td>0.125</td>
</tr>
<tr>
<td>Pre-eclampsia</td>
<td>0.069</td>
</tr>
<tr>
<td>Hypertension, Pregnancy-Induced</td>
<td>0.051</td>
</tr>
<tr>
<td>Hypertension, Renovascular</td>
<td>0.037</td>
</tr>
<tr>
<td>Familial Primary Pulmonary Hypertension</td>
<td>0.015</td>
</tr>
<tr>
<td>Hypertension, Pulmonary</td>
<td>0.008</td>
</tr>
</tbody>
</table>
LIFT GIVES A SENSE OF RELATIONSHIP STRENGTH

\[ \text{lift}(A, B, C) = \frac{P(A \cap B \cap C)}{P(A)P(B)P(C)} \]

High lift indicates good chance of a relationship

\( \text{Lift}((\text{Acquired coagulation factor deficiency, misc. coagulation modifiers})) = 2260.512 \)
\( \text{Lift}((\text{Disease of capillaries, plasma expanders})) = 3273.845 \)

Low lift indicates small chance of relationship

\( \text{Lift}((\text{Diabetes mellitus, antibiotics})) = 0.022 \)
\( \text{Lift}((\text{Neoplasm of uncertain behavior, misc. cardiovascular agents})) = 0.025 \)
## Candidate Phenotype

 Disorders of fluid, electrolyte, and acid-base balance  
 Other and unspecified anemia  
 Antineoplastic agents  
 Antibiotics  
 Calcium channel blocking agents  
 Secondary hypertension  
 Selective immunosuppressants  
 Angiotensin-converting enzyme inhibitors

This candidate phenotype has an average standard deviation above the median of 0.2204.

## Table of Evidence

<table>
<thead>
<tr>
<th>Index</th>
<th>Paper</th>
<th>Standard Deviation above Median Lift</th>
<th>Co-occurrence Tuples</th>
</tr>
</thead>
</table>
| 0     | Title: Currant medical needs in lupus nephritis: solutions through evidence-based, personalized medicine  
Author: Anson, Hana; Joachim; Wiederbusch, Marc; Roen, Brad  
Year: 2015  
View Abstract | 0.051 | (calcium channel blocking agents, selective immunosuppressants) |
| 1     | Title: Assessment of the effects of Low-Level Laser Therapy on the Thyroid Vasculature of Patients with Autoimmune Hypothyroidism by Color Doppler Ultrasound  
Author: Hingbo, Dieno; Boscaini; Chevrette, Maria Cristina; Juliano, Adele G.; Ceri, Giovanni G.; Knob, Mayer; Yatimova, Elisabet M.; Chammes, Matte Cristiano  
Year: 2012  
View Abstract | 0.0196 | (calcium channel blocking agents, selective immunosuppressants) |
| 10    | Title: Fluid and Electrolyte Disturbances in Critically Ill Patients  
Author: Lee, Jay Wook  
Year: 2012  
View Abstract | 0.0004 | (Disorders of fluid, electrolyte, and acid-base balance, hypertension, secondary hypertension) |
| 11    | Title: The Effects of Circadian and Sodium on Blood Pressure in Pediatric Patients with Juvenile Idiopathic Arthritis  
Author: Patel, B. Berger, M. Basha Brown, P. Ingo, D. Nishimura, RW; Zieml, L  
Year: 2012  
View Abstract | 0.0001 | (hypertension, salbutamol, secondary hypertension) |
WEB INTERFACE

CANDIDATE PHENOTYPE

Candidate Phenotype
- Disorders of fluid, electrolyte, and acid-base balance
- Other and unspecified anemia
- Antithrombotic agents
- Calcium channel blocking agents
- Secondary hypertension
- Selective immunosuppressants
- Angiotensin-converting enzyme inhibitors
- Hypertension

This candidate phenotype has an average standard deviation above the median of 0.3064.

Table of Evidence

<table>
<thead>
<tr>
<th>Index</th>
<th>Paper</th>
<th>Standard Deviation above Median Lift</th>
<th>Co-occurrence Tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Title: Ulcer medical needs in lupus erythematosus through evidence-based, personalized medicine&lt;br&gt;Author: Andre, Hans-Joachim; Weber-Busch, Marc; Roos, Brads&lt;br&gt;Year: 2015</td>
<td>0.301</td>
<td>(Calcium channel blocking agents, selective immuno-suppressants)</td>
</tr>
<tr>
<td>1</td>
<td>Title: Assessment of the Effects of Low-Level Laser Therapy on the Thyroid Vasculature of Patients with Autoimmune Hyperthyroidism by Color Doppler Ultrasound&lt;br&gt;Author: Hiltig, Danilo; Borchard, Christian; Maria Greiner; Juliano, Andrea G.; Cerci, Giovanni C.; Knobla, Mayer; Nolte, Elisabeth M.; Chammas, Marla Cristina&lt;br&gt;Year: 2013</td>
<td>0.3195</td>
<td>(Antithrombotic agents, selective immuno-suppressants)</td>
</tr>
<tr>
<td>10</td>
<td>Title: Fluid and Electrolyte Disturbances in Critically Ill Patients&lt;br&gt;Author: Lee, Jay Wook&lt;br&gt;Year: 2013</td>
<td>0.3004</td>
<td>(Disorders of fluid, electrolyte, and acid-base balance, hypertension, secondary hypertension)</td>
</tr>
<tr>
<td>11</td>
<td>Title: The Effects of Catechol or Naproxen on Blood Pressure in Pediatric Patients with Juvenile Idiopathic Arthritis&lt;br&gt;Author: Follmer, B.; Berger, M.; Basko Brown, P.; Ingo, D.; Nitschke, R.W.; Zemel, L.&lt;br&gt;Year: 2013</td>
<td>0.3001</td>
<td>(Hypertension, salicylates, secondary hypertension)</td>
</tr>
</tbody>
</table>
## Candidate Phenotype

### Impact of Article on the Final Score

This candidate phenotype has an average standard deviation (above the median) lift of 0.0004.

### Table of Evidence

<table>
<thead>
<tr>
<th>Index</th>
<th>Paper</th>
<th>Standard Deviation above Median Lifet</th>
<th>Co-occurrence Tuples</th>
</tr>
</thead>
</table>
| 0     | Title: Uterine medical needs in lupus reimplant solutions through evidence-based, personalized medicine  
Author: Andre, Herno-Joachim; Wiederbusch, Marc; Roiv, Brad  
Year: 2015 | 0.001 | (calcium channel blocking agents, selective immunosuppressants) |
| 1     | Title: Assessment of the Effects of Low-Level Laser Therapy on the Thyroid Vascularity of Patients with Autoimmune Hypothyroidism by Color Doppler Ultrasound  
Author: Hingst, Davide; Bandini, Chiara; Girelli, Marco; Grandi, Luciana; Azevedo, G. Gerr; Giovanni G.; Knobel, Mayer; Reinhart, Stefania; Chaminas, Maria Cristina  
Year: 2013 | 0.0196 | (antiplatelet agents, selective immunosuppressants) |
| 10    | Title: Fluid and Electrolyte Disturbances in Critically III Patients  
Author: Lee, Jay Wook  
Year: 2013 | 0.0004 | (Disorders of fluid, electrolyte, and acid-base balance, hypertension, secondary hypertension) |
| 11    | Title: The Effects of Carotid or Neck Pain on Blood Pressure in Pediatric Patients with Juvenile Idiopathic Arthritis  
Author: Folkins, B; Berger, M; Bhattacharay, P; Ingers, D; Nihon, RW; Zimel, L  
Year: 2013 | 0.0001 | (hypertension, statin levels, secondary hypertension) |
WEB INTERFACE

CANDIDATE PHENOTYPE

IMPACT OF ARTICLE ON THE FINAL SCORE

TOP-RANKED ARTICLE RELATED TO THE PHENOTYPE WITH THE ABSTRACT AND LINK TO THE PAPER
## Potential Validation Aid

<table>
<thead>
<tr>
<th>Diagnoses</th>
<th>Medications</th>
<th>Annotator Comment</th>
<th>Score</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypotension, heart failure, cardiac dysrhythmias, unspecified chest pain, ischemic heart disease, hypertension, cardiomyopathy</td>
<td>Statins, proton pump inhibitors, gabapentin, noncardioselective beta blockers, sodium, group v antiarrhythmics, potassium-sparing diuretics</td>
<td>The arrhythmic heart patient</td>
<td>1</td>
<td>317.38</td>
</tr>
<tr>
<td>Disorders of fluid, electrolyte, and acid-base balance; other diseases of lung; hypotension; pleurisy, atelectasis, and pulmonary collapse; unspecified chest pain; other disorders of the kidney and ureter</td>
<td>Anticholinergic bronchodilators, loop diuretics</td>
<td>Lung diseases?</td>
<td>0.417</td>
<td>0.509</td>
</tr>
</tbody>
</table>
SUMMARY

- Produces evidence sets for a candidate phenotype by searching PubMed
- Predicts whether the phenotype is clinically relevant
- Scales to index the entire PubMed Open Access corpus (1M+ articles) using a relatively small machine (i.e., 8 GB RAM, 3 AMD A6-5200 APU with Radeon (TM) HD Graphics processors)
SUMMARY

- Produces evidence sets for a candidate phenotype by searching PubMed
- Predicts whether the phenotype is clinically relevant
- Scales to index the entire PubMed Open Access corpus (1M+ articles) using a relatively small machine (i.e., 8 GB RAM, 3 AMD A6-5200 APU with Radeon (TM) HD Graphics processors)

SYNONYM GENERATION RELIES ON HAND-CURATED ONTOLOGIES — CAN WE DO BETTER?
WORD EMBEDDINGS TO THE RESCUE?

Male-Female
PMCVEC

WORD EMBEDDINGS TO THE RESCUE?

model.most_similar('hypertension', topn=10)

[(u'hypertensive', 0.7680537104606628),
 (u'hypertensions', 0.737869143486023),
 (u'hypertensives', 0.6623083353042603),
 (u'renovascular', 0.642657995223999),
 (u'cardiovascular', 0.6405138969421387),
 (u'antihypertensive', 0.640025138549805),
 (u'prehypertension', 0.635486900806427),
 (u'prehypertensive', 0.6243141293525696),
 (u'rhtn', 0.6119652986526489),
 (u'htn', 0.6072362661361694)]

model.most_similar('diabetes')

[(u'mellitus', 0.9089211225509644),
 (u'prediabetes', 0.7236781120300293),
 (u'diabetic', 0.712675929069519),
 (u'tdm', 0.694568932056427),
 (u'prediabetic', 0.6702020168304443),
 (u'niddm', 0.6690848469734192),
 (u'prediabetics', 0.6682088971138),
 (u'noninsulin', 0.6619926691055298),
 (u'macrovascular', 0.6564791202545166),
 (u'dyslipidemia', 0.6325562000274658)]
Proliferating cell nuclear antigen (PCNA) of formalin-fixed, paraffin-embedded bladder cancer sections was identified by immunohistochemistry in bladder cancer patients. In patients, the cancer cells showed a strong immunoreactivity for PCNA.

Proliferating cell nuclear antigen (PCNA) of formalin-fixed, paraffin-embedded bladder cancer section was identified by immunohistochemistry in bladder cancer patients. In patients, the cancer cells showed a strong immunoreactivity for PCNA.
Proliferating cell nuclear antigen (PCNA) of formalin-fixed, paraffin-embedded bladder cancer section was identified by immunohistochemistry in bladder cancer patients. In patients, the cancer cells showed a strong immunoreactivity for PCNA.
Proliferating cell nuclear antigen (PCNA) of formalin-fixed, paraffin-embedded bladder cancer section was identified by immunohistochemistry in bladder cancer patients. In patients, the cancer cells showed a strong immunoreactivity for PCNA.

RANKING PHRASES USING OUR NEW CRITERIA

Gero, Z, & Ho, J.C., PMVec: Distributed Phrase Representation for Biomedical Text Processing. Journal of Biomedical Informatics. 2019(3):100047
### NEW PHRASE RANKING CRITERIA: INFO FREQ

<table>
<thead>
<tr>
<th>Frequency</th>
<th>JC</th>
</tr>
</thead>
<tbody>
<tr>
<td>present study</td>
<td>stainless steel</td>
</tr>
<tr>
<td>risk factor</td>
<td>myasthenia gravis</td>
</tr>
<tr>
<td>significant difference</td>
<td>endoplasmic reticulum</td>
</tr>
<tr>
<td>cell line</td>
<td>anorexia nervosa</td>
</tr>
<tr>
<td>results suggest</td>
<td>mycophenolate mofetil</td>
</tr>
<tr>
<td>control group</td>
<td>rainbow trout</td>
</tr>
<tr>
<td>amino acid</td>
<td>confidence interval</td>
</tr>
<tr>
<td>significantly high</td>
<td>neurofibrillary tangles</td>
</tr>
<tr>
<td>significantly higher</td>
<td>lupus erythematosus</td>
</tr>
<tr>
<td>risk factors</td>
<td>vena cava</td>
</tr>
</tbody>
</table>
# NEW PHRASE RANKING CRITERIA: INFO FREQ

<table>
<thead>
<tr>
<th><strong>Frequency</strong></th>
<th><strong>JC</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>present study</td>
<td>stainless steel</td>
</tr>
<tr>
<td>risk factor</td>
<td>myasthenia gravis</td>
</tr>
<tr>
<td>significant difference</td>
<td>endoplasmic reticulum</td>
</tr>
<tr>
<td>cell line</td>
<td>anorexia nervosa</td>
</tr>
<tr>
<td>results suggest</td>
<td>mycophenolate mofetil</td>
</tr>
<tr>
<td>control group</td>
<td>rainbow trout</td>
</tr>
<tr>
<td>amino acid</td>
<td>confidence interval</td>
</tr>
<tr>
<td>significantly high</td>
<td>neurofibrillary tangles</td>
</tr>
<tr>
<td>significantly higher</td>
<td>lupus erythematosus</td>
</tr>
<tr>
<td>risk factors</td>
<td>vena cava</td>
</tr>
</tbody>
</table>

**SHORT PHRASES THAT AREN’T ENTIRELY MEDICALLY RELATED**
### NEW PHRASE RANKING CRITERIA: INFO FREQ

<table>
<thead>
<tr>
<th>Frequency</th>
<th>PMI</th>
<th>JC</th>
<th>Word2Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>present study</td>
<td>gemtuzumab ozogamicin</td>
<td>stainless steel</td>
<td>colorectal cancer</td>
</tr>
<tr>
<td>risk factor</td>
<td>erector spinae</td>
<td>myasthenia gravis</td>
<td>waiting list</td>
</tr>
<tr>
<td>significant difference</td>
<td>oculocutaneous albinism</td>
<td>endoplasmic reticulum</td>
<td>virtual screening</td>
</tr>
<tr>
<td>cell line</td>
<td>hpv dna testing</td>
<td>anorexia nervosa</td>
<td>tumor necrosis factor</td>
</tr>
<tr>
<td>results suggest</td>
<td>enterobius vermicularis</td>
<td>mycophenolate mofetil</td>
<td>sodium nitroprusside</td>
</tr>
<tr>
<td>control group</td>
<td>cerebrotendinous xanthomatosis</td>
<td>rainbow trout</td>
<td>sensorineural hearing loss</td>
</tr>
<tr>
<td>amino acid</td>
<td>labrador retrievers</td>
<td>confidence interval</td>
<td>pulmonary arterial hypertension</td>
</tr>
<tr>
<td>significantly high</td>
<td>polymyalgia rheumatica</td>
<td>neurofibrillary tangles</td>
<td>microscopic examination</td>
</tr>
<tr>
<td>significantly higher</td>
<td>lymphomatoid papulosis</td>
<td>lupus erythematosus</td>
<td>glucocorticoid receptor</td>
</tr>
<tr>
<td>risk factors</td>
<td>planum temporale</td>
<td>vena cava</td>
<td>gastric bypass</td>
</tr>
</tbody>
</table>
### NEW PHRASE RANKING CRITERIA: INFO FREQ

<table>
<thead>
<tr>
<th>Frequency</th>
<th>PMI</th>
<th>JC</th>
</tr>
</thead>
<tbody>
<tr>
<td>present study</td>
<td>gemtuzumab ozogamicin</td>
<td>stainless steel</td>
</tr>
<tr>
<td>risk factor</td>
<td>erector spine</td>
<td>myasthenia gravis</td>
</tr>
<tr>
<td>significant difference cell line</td>
<td>oculocutaneous albinism</td>
<td>endoplasmic reticulum</td>
</tr>
<tr>
<td>results suggest</td>
<td>hpv dna testing</td>
<td>anorexia nervosa</td>
</tr>
<tr>
<td>control group</td>
<td>enterobius vermicularis</td>
<td>mycophenolate mofetil</td>
</tr>
<tr>
<td>amino acid</td>
<td>cerebrotendinous xanthomatosis</td>
<td>rainbow trout</td>
</tr>
<tr>
<td>significantly high risk factors</td>
<td>labrador retrievers</td>
<td>confidence interval</td>
</tr>
<tr>
<td>significantly higher risk factors</td>
<td>polymyalgia rheumatica</td>
<td>neurofibrillary tangles</td>
</tr>
<tr>
<td></td>
<td>lymphomatoid papulosis</td>
<td>lupus erythematosus</td>
</tr>
<tr>
<td></td>
<td>planum temporale</td>
<td>vena cava</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word2Phrase</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>colorectal cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>waiting list</td>
<td></td>
<td></td>
</tr>
<tr>
<td>virtual screening</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tumor necrosis factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sodium nitroprusside</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensorineural hearing loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pulmonary arterial hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>microscopic examination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>glucocorticoid receptor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gastric bypass</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SLIGHTLY MORE MEDICALLY-RELATED TERMS BUT STILL SHORT**
### NEW PHRASE RANKING CRITERIA: INFO FREQ

<table>
<thead>
<tr>
<th>Frequency</th>
<th>PMI</th>
<th>JC</th>
<th>Word2Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>present study</td>
<td>gemtuzumab ozogamicin</td>
<td>stainless steel</td>
<td>colorectal cancer</td>
</tr>
<tr>
<td>risk factor</td>
<td>erector spinae</td>
<td>myasthenia gravis</td>
<td>waiting list</td>
</tr>
<tr>
<td>significant difference</td>
<td>oculocutaneous albinism</td>
<td>endoplasmic reticulum</td>
<td>virtual screening</td>
</tr>
<tr>
<td>cell line</td>
<td>hpv dna testing</td>
<td>anorexia nervosa</td>
<td>tumor necrosis factor</td>
</tr>
<tr>
<td>results suggest</td>
<td>enterobius vermicularis</td>
<td>mycophenolate mofetil</td>
<td>sodium nitroprusside</td>
</tr>
<tr>
<td>control group</td>
<td>cerebrotendinous xanthomatosis</td>
<td>rainbow trout</td>
<td>sensorineural hearing loss</td>
</tr>
<tr>
<td>amino acid</td>
<td>labrador retrievers</td>
<td>confidence interval</td>
<td>pulmonary arterial hypertension</td>
</tr>
<tr>
<td>significantly high</td>
<td>polymyalgia rheumatica</td>
<td>neurofibrillary tangles</td>
<td>microscopic examination</td>
</tr>
<tr>
<td>significantly higher</td>
<td>lymphomatoid papulosis</td>
<td>lupus erythematosus</td>
<td>glucocorticoid receptor</td>
</tr>
<tr>
<td>risk factors</td>
<td>planum temporale</td>
<td>vena cava</td>
<td>gastric bypass</td>
</tr>
</tbody>
</table>


### New Phrase Ranking Criteria: Info Freq

<table>
<thead>
<tr>
<th>Frequency</th>
<th>PMI</th>
<th>JC</th>
<th>Info_Freq</th>
<th>Word2Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>present study</td>
<td>gemtuzumab ozogamicin</td>
<td>stainless steel</td>
<td>polymerase chain reaction</td>
<td>colorectal cancer</td>
</tr>
<tr>
<td>risk factor</td>
<td>erector spine</td>
<td>myasthenia gravis</td>
<td>magnetic resonance imaging</td>
<td>waiting list</td>
</tr>
<tr>
<td>significant difference</td>
<td>oculocutaneous albinism</td>
<td>endoplasmic reticulum</td>
<td>vascular endothelial growth factor</td>
<td>virtual screening</td>
</tr>
<tr>
<td>cell line</td>
<td>hpv dna testing</td>
<td>anorexia nervosa</td>
<td>chronic obstructive pulmonary disease</td>
<td>tumor necrosis factor</td>
</tr>
<tr>
<td>results suggest</td>
<td>enterobius vermicularis</td>
<td>mycophenolate mofetil</td>
<td>coronary artery bypass graft</td>
<td>sodium nitroprusside</td>
</tr>
<tr>
<td>control group</td>
<td>cerebrotendinous xanthomatosis</td>
<td>rainbow trout</td>
<td>receiver operating characteristic curve</td>
<td>sensorineural hearing loss</td>
</tr>
<tr>
<td>amino acid</td>
<td>labrador retrievers</td>
<td>confidence interval</td>
<td>body mass index</td>
<td>pulmonary arterial hypertension</td>
</tr>
<tr>
<td>significantly high</td>
<td>polymyalgia rheumatic</td>
<td>neurofibrillary tangles</td>
<td>reverse transcription polymerase chain reaction</td>
<td>microscopic examination</td>
</tr>
<tr>
<td>significantly higher</td>
<td>lymphomatoid papulosis</td>
<td>lupus erythematosus</td>
<td>left anterior descending coronary artery</td>
<td>glucocorticoid receptor</td>
</tr>
<tr>
<td>risk factors</td>
<td>planum temporale</td>
<td>vena cava</td>
<td>amino acid</td>
<td>gastric bypass</td>
</tr>
</tbody>
</table>
# New Phrase Ranking Criteria: Info Freq

<table>
<thead>
<tr>
<th>Frequency</th>
<th>PMI</th>
<th>JC</th>
<th>Info_Freq</th>
<th>Word2Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>present study</td>
<td>gemtuzumab ozogamicin</td>
<td>stainless steel</td>
<td>polymerase chain reaction</td>
<td>colorectal cancer</td>
</tr>
<tr>
<td>risk factor</td>
<td>erector spinae</td>
<td>myasthenia gravis</td>
<td>magnetic resonance imaging</td>
<td>waiting list</td>
</tr>
<tr>
<td>significant difference</td>
<td>oculocutaneous albinism</td>
<td>endoplasmic reticulum</td>
<td>vascular endothelial growth factor</td>
<td>virtual screening</td>
</tr>
<tr>
<td>cell line</td>
<td>hpv dna testing</td>
<td>anorexia nervosa</td>
<td>chronic obstructive pulmonary disease</td>
<td>tumor necrosis factor</td>
</tr>
<tr>
<td>results suggest</td>
<td>enterobius vermicularis</td>
<td>mycophenolate mofetil</td>
<td>coronary artery bypass graft</td>
<td>sodium nitroprusside</td>
</tr>
<tr>
<td>control group</td>
<td>cerebrotendinous xanthomatosis</td>
<td>rainbow trout</td>
<td>receiver operating characteristic curve</td>
<td>sensorineural hearing loss</td>
</tr>
<tr>
<td>amino acid</td>
<td>labrador retrievers</td>
<td>confidence interval</td>
<td>body mass index</td>
<td>pulmonary arterial hypertension</td>
</tr>
<tr>
<td>significantly high</td>
<td>polymyalgia rheumatic</td>
<td>neurofibrillary tangles</td>
<td>reverse transcription polymerase chain reaction</td>
<td>microscopic examination</td>
</tr>
<tr>
<td>significantly higher</td>
<td>lymphomatoid papulosis</td>
<td>lupus erythematosus</td>
<td>left anterior descending coronary artery</td>
<td>glucocorticoid receptor</td>
</tr>
<tr>
<td>risk factors</td>
<td>planum temporale</td>
<td>vena cava</td>
<td>amino acid</td>
<td>gastric bypass</td>
</tr>
</tbody>
</table>

**Good Balance of Phrase Length, Frequency and Medical Relevance**
QUANTITATIVE COMPARISON: BIOMEDICAL SIMILARITY TASKS
QUANTITATIVE COMPARISON: BIOMEDICAL SIMILARITY TASKS

HIGHER IS BETTER
QUANTITATIVE COMPARISON: BIOMEDICAL SIMILARITY TASKS

HIGHER IS BETTER

OUR MODEL
QUALITATIVE COMPARISON

(a) No Phrases
(b) PubMed Phrases
(c) PMCVec
SUMMARY

- Learns quality vector embeddings for single word and multi-word phrases
- Generates useful multi-word phrases automatically from the corpus
- Can be widely used for a variety of biomedical-NLP tasks
SUMMARY

- Learns quality vector embeddings for single word and multi-word phrases
- Generates useful multi-word phrases automatically from the corpus
- Can be widely used for a variety of biomedical-NLP tasks

CAN WE LEVERAGE PMCVEC TO SUMMARIZE ARTICLES SUCCINCTLY?
Application of next generation sequencing in clinical microbiology and infection prevention

Current molecular diagnostics of human pathogens provide limited information that is often not sufficient for outbreak and transmission investigation. Next generation sequencing (NGS) determines the DNA sequence of a complete bacterial genome in a single sequence run, and from these data, information on resistance and virulence, as well as information for typing is obtained, useful for outbreak investigation. The obtained genome data can be further used for the development of an outbreak-specific screening test. In this review, a general introduction to NGS is presented, including the library preparation and the major characteristics of the most common NGS platforms, such as the MiSeq (Illumina) and the Ion PGM™ (ThermoFisher). An overview of the software used for NGS data analyses used at the medical microbiology diagnostic laboratory in the University Medical Center Groningen in The Netherlands is given. Furthermore, applications of NGS in the clinical setting are described, such as outbreak management, molecular case finding, characterization and surveillance of pathogens, rapid identification of bacteria using the 16S-23S rRNA region, taxonomy, metagenomics approaches on clinical samples, and the determination of the transmission of zoonotic micro-organisms from animals to humans. Finally, we share our vision on the use of NGS in personalised microbiology in the near future, pointing out specific requirements.
Application of next generation sequencing in clinical microbiology and infection prevention

Current molecular diagnostics of human pathogens provide limited information that is often not sufficient for outbreak and transmission investigation. Next generation sequencing (NGS) determines the DNA sequence of a complete bacterial genome in a single sequence run, and from these data, information on resistance and virulence, as well as information for typing is obtained, useful for outbreak investigation. The obtained genome data can be further used for the development of an outbreak-specific screening test. In this review, a general introduction to NGS is presented, including the library preparation and the major characteristics of the most common NGS platforms, such as the MiSeq (Illumina) and the Ion PGM™ (ThermoFisher). An overview of the software used for NGS data analyses used at the medical microbiology diagnostic laboratory in the University Medical Center Groningen in The Netherlands is given. Furthermore, applications of NGS in the clinical setting are described, such as outbreak management, molecular case finding, characterization and surveillance of pathogens, rapid identification of bacteria using the 16S-23S rRNA region, taxonomy, metagenomics approaches on clinical samples, and the determination of the transmission of zoonotic micro-organisms from animals to humans. Finally, we share our vision on the use of NGS in personalised microbiology in the near future, pointing out specific requirements.

KeyPhrases:
Clinical microbiology
Infection prevention
Ion PGM
MiSeq
Next generation sequencing
Whole genome sequencing
Application of next generation sequencing in clinical microbiology and infection prevention

Current molecular diagnostics of human pathogens provide limited information that is often not sufficient for outbreak and transmission investigation. Next generation sequencing (NGS) determines the DNA sequence of a complete bacterial genome in a single sequence run, and from these data, information on resistance and virulence, as well as information for typing is obtained, useful for outbreak investigation. The obtained genome data can be further used for the development of an outbreak-specific screening test. In this review, a general introduction to NGS is presented, including the library preparation and the major characteristics of the most common NGS platforms, such as the MiSeq (Illumina) and the Ion PGM™ (ThermoFisher). An overview of the software used for NGS data analyses used at the medical microbiology diagnostic laboratory in the University Medical Center Groningen in The Netherlands is given. Furthermore, applications of NGS in the clinical setting are described, such as outbreak management, molecular case finding, characterization and surveillance of pathogens, rapid identification of bacteria using the 16S-23S rRNA region, taxonomy, metagenomics approaches on clinical samples, and the determination of the transmission of zoonotic micro-organisms from animals to humans. Finally, we share our vision on the use of NGS in personalised microbiology in the near future, pointing out specific requirements.

KeyPhrases:
- Clinical microbiology
- Infection prevention
- Ion PGM
- MiSeq
- Next generation sequencing
- Whole genome sequencing
Application of next generation sequencing in clinical microbiology and infection prevention

Current molecular diagnostics of human pathogens provide limited information that is often not sufficient for outbreak and transmission investigation. Next generation sequencing (NGS) determines the DNA sequence of a complete bacterial genome in a single sequence run, and from these data, information on resistance and virulence, as well as information for typing is obtained, useful for outbreak investigation. The obtained genome data can be further used for the development of an outbreak-specific screening test. In this review, a general introduction to NGS is presented, including the library preparation and the major characteristics of the most common NGS platforms, such as the MiSeq (Illumina) and the Ion PGM™ (ThermoFisher). An overview of the software used for NGS data analyses used at the medical microbiology diagnostic laboratory in the University Medical Center Groningen in The Netherlands is given. Furthermore, applications of NGS in the clinical setting are described, such as outbreak management, molecular case finding, characterization and surveillance of pathogens, rapid identification of bacteria using the 16S-23S rRNA region, taxonomy, metagenomics approaches on clinical samples, and the determination of the transmission of zoonotic micro-organisms from animals to humans. Finally, we share our vision on the use of NGS in personalised microbiology in the near future, pointing out specific requirements.
Application of next generation sequencing in clinical microbiology and infection prevention

Current molecular diagnostics of human pathogens provide limited information that is often not sufficient for outbreak and transmission investigation. Next generation sequencing (NGS) determines the DNA sequence of a complete bacterial genome in a single sequence run, and from these data, information on resistance and virulence, as well as information for typing is obtained, useful for outbreak investigation. The obtained genome data can be further used for the development of an outbreak-specific screening test. In this review, a general introduction to NGS is presented, including the library preparation and the major characteristics of the most common NGS platforms, such as the MiSeq (Illumina) and the Ion PGM™ (ThermoFisher). An overview of the software used for NGS data analyses used at the medical microbiology diagnostic laboratory in the University Medical Center Groningen in The Netherlands is given. Furthermore, applications of NGS in the clinical setting are described, such as outbreak management, molecular case finding, characterization and surveillance of pathogens, rapid identification of bacteria using the 16S-23S rRNA region, taxonomy, metagenomics approaches on clinical samples, and the determination of the transmission of zoonotic micro-organisms from animals to humans. Finally, we share our vision on the use of NGS in personalised microbiology in the near future, pointing out specific requirements.
Application of next generation sequencing in clinical microbiology and infection prevention

Current molecular diagnostics of human pathogens provide limited information that is often not sufficient for outbreak and transmission investigation. Next generation sequencing (NGS) determines the DNA sequence of a complete bacterial genome in a single sequence run, and from these data, information on resistance and virulence, as well as information for typing is obtained, useful for outbreak investigation. The obtained genome data can be further used for the development of an outbreak-specific screening test. In this review, a general introduction to NGS is presented, including the library preparation and the major characteristics of the most common NGS platforms, such as the MiSeq (Illumina) and the Ion PGM™ (ThermoFisher). An overview of the software used for NGS data analyses used at the medical microbiology diagnostic laboratory in the University Medical Center Groningen in The Netherlands is given. Furthermore, applications of NGS in the clinical setting are described, such as outbreak management, molecular case finding, characterization and surveillance of pathogens, rapid identification of bacteria using the 16S-23S rRNA region, taxonomy, metagenomics approaches on clinical samples, and the determination of the transmission of zoonotic micro-organisms from animals to humans. Finally, we share our vision on the use of NGS in personalised microbiology in the near future, pointing out specific requirements.
Application of next generation sequencing in clinical microbiology and infection prevention

Current molecular diagnostics of human pathogens provide limited information that is often not sufficient for outbreak and transmission investigation. Next generation sequencing (NGS) determines the DNA sequence of a complete bacterial genome in a single sequence run, and from these data, information on resistance and virulence, as well as information for typing is obtained, useful for outbreak investigation. The obtained genome data can be further used for the development of an outbreak-specific screening test. In this review, a general introduction to NGS is presented, including the library preparation and the major characteristics of the most common NGS platforms, such as the MiSeq (Illumina) and the Ion PGM™ (ThermoFisher). An overview of the software used for NGS data analyses used at the medical microbiology diagnostic laboratory in the University Medical Center Groningen in The Netherlands is given. Furthermore, applications of NGS in the clinical setting are described, such as outbreak management, molecular case finding, characterization and surveillance of pathogens, rapid identification of bacteria using the 16S-23S rRNA region, taxonomy, metagenomics approaches on clinical samples, and the determination of the transmission of zoonotic micro-organisms from animals to humans. Finally, we share our vision on the use of NGS in personalised microbiology in the near future, pointing out specific requirements.
Implantable cardioverter defibrillators in lamin A/C mutation carriers with cardiac conduction disorders. Sudden cardiac death is frequent in patients with lamin A/C gene (LMNA) mutations and may be related to ventricular arrhythmias (VA).
Implantable cardioverter defibrillators in lamin A/C mutation carriers with cardiac conduction disorders. Sudden cardiac death is frequent in patients with lamin A/C gene (LMNA) mutations and may be related to ventricular arrhythmias (VA).
NAMEDKEYS

OVERALL PROCESS

IDENTIFYING POSSIBLE KEYPHRASES

USE WORD EMBEDDINGS AND PHRASE RANKING TO FIND MEANINGFUL KEYPHRASES SIMILAR TO THE DOCUMENT

Abstract
Implantable cardioverter defibrillators in lamin A/C mutation carriers with cardiac conduction disorders. Sudden cardiac death is frequent in patients with lamin A/C gene (LMNA) mutations and may be related to ventricular arrhythmias (VA).

Named Entities
- cardiac conduction disorders
- sudden cardiac death
- lamin A/C gene
- ventricular arrhythmias
- cardiac
- LMNA

words/phrases
- Implantable cardioverter-defibrillators
- lamin A/C mutation
- patients
- mutations
- VA

NER

Chunking

Embed

Idf-weighting

DoC Embed

Candidates

Phrase Quality Score

Cluster

Return top candidates

keyphrases
- sudden cardiac death
- ventricular arrhythmia
- cardiac
- cardiac conduct disorder
- lamin A/C gene
- lmna
- lamin A/C mutation
OVERALL PROCESS

Roger

USE WORD EMBEDDINGS AND PHRASE RANKING TO FIND MEANINGFUL KEYPHRASES SIMILAR TO THE DOCUMENT.

IDENTIFYING POSSIBLE KEYPHRASES

FIND DIVERSE AND REPRESENTATIVE KEYPHRASES

NER

Chunking

Named Entities

words/phrases

Abstract

Implantable cardioverter defibrillators in lamin A/C mutation carriers with cardiac conduction disorders. Sudden cardiac death is frequent in patients with lamin A/C gene (LAMA) mutations and may be related to ventricular arrhythmias (VA).

NER

Chunking

Named Entities

cardiac conduction disorders
sudden cardiac death
lamin A/C gene
ventricular arrhythmias
ventricular
cardiac
LAMA

words/phrases

Implantable cardioverter-defibrillators
lamin A/C mutation
patients
mutations
VA

Idf-weighting

DoC Embed

Phrase Quality Score

Cluster

keyphrases

Sudden cardiac death
ventricular arrhythmia
cardiac
cardiac conduct disorder
lamin A/C
lamin A/C mutation
## PHRASE QUALITY USING INFO_FREQ

<table>
<thead>
<tr>
<th>Keyphrases correctly extracted by NamedKeys:</th>
<th>Common phrases incorrectly extracted by baseline methods:</th>
</tr>
</thead>
<tbody>
<tr>
<td>prostate cancer 0.81</td>
<td>different prostate cancer risk groups 0</td>
</tr>
<tr>
<td>radical prostatectomy 0.84</td>
<td>high risk prostate cancer 0.32</td>
</tr>
<tr>
<td>magnetic resonance imaging 0.92</td>
<td>high risk cohort 0.17</td>
</tr>
<tr>
<td>positive predictive value 0.93</td>
<td>high risk categories 0.12</td>
</tr>
<tr>
<td>negative predictive value 0.93</td>
<td>tesla multiparametric magnetic resonance imaging 0</td>
</tr>
<tr>
<td>multivariate analyses 0.91</td>
<td>low risk subjects 0.1</td>
</tr>
<tr>
<td>endorectal coil 0.89</td>
<td>low risk cohort 0.1</td>
</tr>
<tr>
<td>extraprostatic extension 0.75</td>
<td>multiparametric magnetic resonance imaging 0.25</td>
</tr>
<tr>
<td></td>
<td>multiparametric magnetic 0</td>
</tr>
<tr>
<td></td>
<td>determined sensitivity 0</td>
</tr>
</tbody>
</table>


Repetitive methylene blue-mediated photoantimicrobial chemotherapy changes the susceptibility and expression of the outer membrane proteins of Pseudomonas aeruginosa. Since bacterial multidrug efflux pumps mediate intracellular photosensitizer methylene blue, a change in the expression alters the susceptibility to photoantimicrobial chemotherapy (PACT) of Pseudomonas aeruginosa, which may occur following repetitive sublethal challenges.
**EXPERIMENTAL RESULTS**

- New benchmark dataset
- 3049 PubMed Open Access articles
  - Abstracts with at least 5 author-provided keyphrases
  - Title of the article
  - Abstract of the article
  - List of keyphrases provided by the authors
SUMMARY

- Ability to identify diverse and representative keyphrases
- Potentially improved document representation using keyphrases
- Creation of a new benchmark dataset for biomedical keyphrase extraction
CONCLUSIONS

- Pheknow-Cloud/PIVET: Tool to produce evidence sets for a candidate phenotype by searching PubMed
- PMCVec: Quality vector embeddings for single word and multi-word phrases
- NamedKeys: Keyphrase summarization of PubMed articles
- Future directions
  - How to do generate implicit keyphrases (phrases not appearing in the corpus)?
  - Can these ideas be used to automate systematic review process?
ACKNOWLEDGEMENTS

COLLABORATORS

- Byron Wallace (Northeastern)
- Jette Henderson (Cognitive Scale)
- Joydeep Ghosh (UT Austin)
- Junyuan Ke (University of Rochester)
- Ryan Bridges (Epic Systems)
- Zelalem Gero (Emory)