Programa Doutoral em Ciência de Dados de Saúde

Faculdade de Medicina da Universidade do Porto

Projeto Laboratorial

RELATÓRIO FINAL

GRUPO MIDAS

Daniel Magano

Joana Moreira

Paulo Dias Costa

Pedro Amorim

Porto, Maio de 2022

1.Introdução

A Ciência, nas suas múltiplas dimensões, desempenha um papel cada vez mais importante na sociedade moderna. No contexto clínico este papel é ainda mais relevante na medida em que se podem aplicar os resultados da investigação à própria população de estudo, com tradução direta na melhoria da prestação de cuidados e na definição de novas políticas de saúde. Os dados colhidos na prática clínica quotidiana geram informação, podendo esta ser utilizada para produção de conhecimento quando aplicada na investigação. O conhecimento produzido neste contexto suporta a tomada de decisão, baseada em evidência, auxiliando assim a prática clínica. Por sua vez, esta "nova" prática gera dados e informação adicional, mantendo-se assim um ciclo potencialmente perpétuo de geração de conhecimento que melhora continuamente o processo e, principalmente, a forma como se faz medicina/saúde e se prestam cuidados à população.

Em Portugal, temos assistido nos últimos anos a um crescimento substancial da investigação em saúde, com reflexo no aumento do número de publicações e no impacto científico das mesmas. Para além de um sinal de maturidade da investigação clínica, estes dados expressam também a cada vez maior diferenciação dos profissionais de saúde no nosso país e afirmam a medicina baseada na evidência e a figura do profissional de saúde/cientista como partes fundamentais da saúde no nosso país. De facto, a medicina baseada na evidência - enquanto prática clínica assente na melhor e mais recente evidência científica disponível - carece de investigação de qualidade, atual e passível de ser aplicada à população com a qual cada um trabalha. Isso exige que os profissionais de saúde saibam pesquisar, interpretar e aplicar essa evidência de acordo com as necessidades de cada paciente. Também por estes motivos, existem atualmente cláusulas nos próprios contratos-programa dos hospitais, que incentivem à investigação clínica. Apesar de o financiamento dedicada a esta matéria ser percentualmente baixo, a verdade é que é também uma forma de incentivar as unidades hospitalares à investigação e à participação em ensaios clínicos.

Num momento em que se faz cada vez mais investigação em contexto hospitalar e que se prevê que esta cresça ainda mais, é difícil aceder a estas métricas, não existindo um repositório que permita consultar de forma unificada e centralizada a produção científica dos hospitais portugueses. Esta informação pode ser útil não só aos próprios hospitais mas também aos profissionais de saúde, sendo importante aceder de forma fácil a esses dados para que seja possível perceber a dimensão e evolução da produção científica hospitalar, as áreas de investigação em que se focam e o contexto em que essa investigação é realizada. Assim, o nosso objetivo primário é criar uma plataforma que permita consultar a literatura hospitalar realizada nos últimos cinco anos, em Portugal. Como objetivo secundário, pretendemos que esta ferramenta permita a filtragem dos dados por hospital/tipologia de hospital, ano, autor e tipo de estudo realizado.

2.Métodos

Fontes de dados

De forma a obtermos os dados de produção científica dos hospitais portugueses, realizamos uma pesquisa de largo espectro na Medline (através da PubMed[1]) utilizando permutações de "hospital" ou "hospitalar" , combinadas com o termo "Portugal", na afiliação dos autores e utilizando como filtro apenas estudos realizados em humanos entre 2017 e 2021. Para simplificação, a *query* utilizada encontra-se descrita abaixo:

```
(("hospital"[Affiliation] OR "hospitalar"[Affiliation]) AND ("Portugal"[Affiliation])
AND ((humans[Filter]) AND (2017/01/01:2021/12/31[pdat]))
```

Extração de dados

A extração de dados foi efetuada com recurso à *library* 'Metapub'[2] que permite realizar o *download* automatizado de informação da PubMed, utilizando código em *Python* (Anexo I). O código utilizado consiste num pedido de artigo (com campos especificados) efetuado a cada 10 segundos ao servidor da PubMed, limite este imposto pela própria PubMed. O resultado deste pedido resulta num ficheiro em formato .csv que constitui a base da nossa análise.

Do ficheiro resultante foram selecionadas as seguintes variáveis:

- PMID
- URL
- Revista
- Título
- doi
- Ano
- majorMESH (MESH Unique ID)
- Outros MESH (MESH Unique ID)
- Autores
- Afiliação
- Tipo de artigo

Preparação de dados

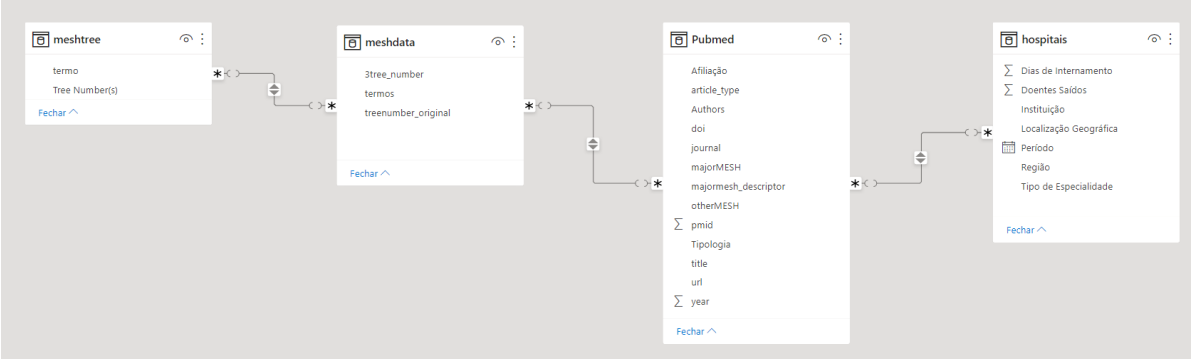
Os dados em .csv foram trabalhados com recurso ao R (R Foundation for Statistical Computing, Vienna, Austria), PowerBI (Microsoft Corporation, Redmond, Washington, USA) e Excel (Microsoft Corporation,

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://pypi.org/project/metapub/

Redmond, Washington, USA). Utilizando as mesmas ferramentas, e tendo como objetivo a classificação da literatura produzida em meio hospitalar, procedemos à remoção das linhas que continham campos em branco nas "afiliações". Estes campos em branco são causados por uma limitação da libraria MetaPub.

Utilizando a ferramenta editor do Power Query do PowerBI foi adicionada uma variável à tabela "pubmed" com o termo MESH correspondente ao código MESH Unique ID utilizando uma API[3] da Pubmed:
Json.Document(Web.Contents("https://id.nlm.nih.gov/mesh/lookup/label?resource="&[majorMESH]))

Foram adicionadas três tabelas relacionais ao modelo, de acordo com o esquema descrito na Figura 1 abaixo.



A Tabela "hospitais" obtida do Portal da transparência[4] contêm as seguintes variáveis:

- Instituição (hospitais portugueses) - *relacionada com a variável Afiliação da tabela "pubmed";
- Localização geográfica (latitude e longitude);
- Região;
- Dias de internamento (Número de dias de internamento utilizados por todos os doentes internados num período, cuja permanência de internamento seja superior a 24h, excluindo o dia da alta);
- Doentes saídos (Indicador que mede a produção em internamento considerando todos os doentes que têm alta do internamento de um estabelecimento de saúde num período de referência);
- Período;
- Tipo de especialidade.

A Tabela "meshtree"[5] obtida do site da Pubmed contém as seguintes variáveis:

- Tree Number(s) *relacionada com variável 3tree_number da tabela "meshdata";
- Termo.

[3] https://id.nlm.nih.gov/mesh/swagger/ui
[4] https://transparencia.sns.gov.pt/explore/dataset/atividade-de-internamento-hospitalar/information/?disjunctive.regiao&disjunctive.instituicao&disjunctive.tipo_de_especialidade&sort=tempo
[5] https://www.nlm.nih.gov/mesh/2020/download/2020MeShTreeHierarchy.xlsx

A tabela "meshdata" foi criada de novo em PowerBI através da listagem da variável "majormesh_descriptor" (termos mesh unique ID) da tabela "pubmed":

- Termos (termos mesh unique ID)* relacionada com "majormesh_descriptor" da tabela pubmed;
- Treenumber_original ("treenumber" da tabela "meshtree");
- 3tree_number (3 primeiros algarismos da variável Treenumber_original) * relacionada com "treenumber(s) da tabela "meshtree").

Visualização de dados

Para a elaboração do *dashboard* foi utilizada a aplicação Power BI, na medida em que entrega, de forma mais fácil e intuitiva, uma coletânea de serviços de análise de dados que permitem a recolha, tratamento e estruturação de informações em apenas um painel.

O PowerBi permite integrar grandes quantidades de dados, mesmo que provenientes de diferentes bases de dados, numa única plataforma capaz de os agrupar e filtrar, apresentando-os de forma numérica e gráfica, mas mais simples e intuitiva. O objetivo é integrar vários dados e apresentá-los em informações coerentes, interativas e visualmente interessantes.

O acesso ao *dashboard* deve ser feito através da conta institucional da Universidade do Porto, seguindo este link ou acedendo através do QR code abaixo (Figura 2).

Power BI

No primeiro separador - Produção Científica Hospitalar, o utilizador consegue ter uma visão da produção (n) científica dos hospitais portugueses através de um gráfico de barras. Para além do número total de artigos publicados em cada ano, cada barra discrimina também os tipos de artigos incluídos. Na parte inferior do ecrã, o utilizador pode também consultar as listas dos hospitais e dos tipos de artigos, conseguindo ordenar de forma crescente ou decrescente de ocorrência. Finalmente, é possível ainda ao utilizador filtrar por hospital (ou tipo de hospital) e por tipo de artigo.

No segundo separador é apresentado o número de artigos por hospital num mapa onde o círculo que marca cada centro hospitalar com cor diferente aumenta de acordo com o número de publicações. Também neste separador é possível filtrar os dados por região do país, tipologia hospitalar e ano da

publicação. Por fim, um mapa com a mediana de altas hospitalares por mês (como métrica da casuística do hospital) foi também colocado aqui com o objetivo de permitir uma comparação rápida entre a produção científica e a casuística de um dado hospital.

No terceiro separador - Revista - é apresentado um gráfico com o número de publicações em cada revista. Também aqui, o tamanho do retângulo que representa cada revista aumenta de acordo com o número de publicações. Ao clicar em cada revista é possível consultar não só o número de artigos publicados nessa revista, como uma lista (na parte inferior do ecrã) de todos os artigos publicados nessa revista - com título, autores e link para o respetivo artigo. Um filtro foi também colocado neste separador, para que possa ser possível filtrar por hospital e ano de publicação.

No separador “Mesh” é possível efetuar uma pesquisa no título e/ou uma pesquisa por autor. Após a pesquisa, o utilizador consegue ver não só os termos MeSH (e os descritores MeSH) mais comuns numa dada área de interesse, como consultar uma lista dos artigos publicados nessa área durante os últimos 5 anos - com possibilidade de clicar no link para o respetivo artigo. A mesma pesquisa pode ser efetuada através do nome do autor, para ser possível encontrar o trabalho científico realizado por alguém.

3.Resultados

Foi feito o *download* de 5.599 artigos publicados por autores associados a hospitais portugueses. Durante este período, a produção científica hospitalar tem vindo a subir de forma consecutiva - de 2017, com 1683 artigos, até 2021, com 2520 artigos publicados - apenas com um ligeiro abrandamento no ano de 2020 (possivelmente provocado pelo início da pandemia por COVID-19). Os tipos de artigos mais frequentes são: '*Journal Article*' (4837) e os '*Case Reports*' (1300).

Os hospitais com mais produção científica foram: Centro Hospitalar Universitário de São João (1149), Centro Hospitalar Universitário de Coimbra (1041), Centro Hospitalar Universitário do Porto (657) e o Centro Hospitalar Universitário de Lisboa Central (572) e Centro Hospitalar Universitário de Lisboa Norte (521). Já os hospitais distritais com mais produção foram o Hospital de Braga (267) e o Hospital Garcia da Orta (217). Dentro dos hospitais especializados, os mais profícuos foram o Hospital Psiquiátrico de Lisboa (24) e o Hospital Magalhães Lemos (22). Relativamente às Unidades Locais de Saúde (ULS), destacam-se a ULS de Matosinhos (102) e a ULS do Alto Minho (39).

A produção científica dos hospitais portugueses foi publicada durante este período em centenas de revistas diferentes. As mais comuns foram: BMJ Case Reports (358), Acta Médica Portuguesa (356), Revista Portuguesa de Cardiologia (157) e Pulmonology (81).

Por fim, em relação aos MeSH *tree heading*, os mais comuns foram: *Diagnosis* (309), *Surgical Procedures, Operative* (290), *Cardiovascular Diseases* (264) e *Therapeutics* (249). Se nos focarmos nos MeSh *Descriptors,* os termos mais comuns são: *Quality of Life* (45), *Severity of Ilness Index* (30) e *Visual Acuity* (28). Um facto interessante (embora esperado) é o aparecimento das expressões '*Environment and Public Health*' e '*Pandemics'* entre os termos mais comuns no ano de 2020.

4.Discussão/Conclusão

Cumprimento dos objetivos

De acordo com a ideia definida inicialmente, conseguimos cumprir os principais objetivos a que nos propusemos. De forma sumária, conseguimos criar uma plataforma que permite consultar a produção científica realizada por autores com associação a hospitais portugueses nos últimos cinco anos; e que permite, entre outras coisas, filtrar os dados por hospital, por ano, por autor e por tipo de estudo realizado.

Limitações

Existem algumas limitações no nosso trabalho. O primeiro prende-se com o facto das publicações não representarem verdadeiramente a produção científica hospitalar (realizado nos diversos hospitais em Portugal), mas sim produzida por autores com associação a hospitais portugueses.

Outra das limitações é relativa à remoção de publicações em que a informação de afiliações não foi extraída, devido a uma limitação do *MetaPub*, libraria de *Python*. De forma a colmatar esta falha do código utilizado, foi encontrado uma alternativa em R que aparentemente conseguia extrair mais informação relativa a estes dados, embora estes fossem extraídos com uma organização inferior à da obtida através do *Python*. Por limite de tempo, não foi possível explorar a fundo esta alternativa, mas será a melhor solução para o problema decorrente da utilização do *MetaPub*.

Por fim, o *PowerBI* é uma aplicação que, embora seja muito útil, depende sempre de uma licença associada de forma a permitir a disponibilização de conteúdo publicamente.

Trabalho Futuro

Apesar de termos construído uma ferramenta promissora, que permite responder aos nossos principais objetivos (embora reconheçamos as limitações presentes), existem outros passos que pensamos que seriam interessantes realizar no futuro.

A primeira sugestão seria a atualização periódica automática dos dados guardados na(s) base(s) de dados, de forma que o utilizador consultasse informação a mais atualizada quanto possível. Um outro ponto que podia ser interessante seria incluir um período temporal mais alargado, de 10 anos, por exemplo, embora tenhamos notado que um volume adicional de dados impacta a performance da plataforma.

De forma mais abrangente, tentar melhorar a filtragem dos dados de forma a incluir de facto a produção hospitalar e não a produção realizada por profissionais que trabalham em contexto hospitalar seria fundamental para validar a nossa pretensão inicial. Durante o tempo de realização deste trabalho não encontramos uma forma viável de o fazer (sem ter que consultar a metodologia do trabalho) mas o

objetivo mantém-se para futuro. Ainda neste âmbito mais geral, termos uma associação destes dados com o financiamento hospital poderia ser bastante útil.

Dentro da nossa ferramenta no Power BI, seria também útil acrescentar algumas funcionalidades. No separador 'Revista', incluir o quartil da revista como opção de pesquisa; acrescentar um novo separador quer permitisse uma pesquisa direta por termos no *abstract* do artigo; e, finalmente, incluir uma opção de pesquisa por área de estudo / especialidade.

Anexo I - código *Python*

```
import hashlib
import time
import xml.etree.ElementTree as ET
import pandas as pd
from metapub import PubMedFetcher



def props(cls):
    return [i for i in cls._dict.keys() if i[:1] != '']



query = """(("hospital"[Affiliation] OR "hospitalar"[Affiliation]) AND "Portugal"[Affiliation]) AND
((humans[Filter]) AND (2017/01/01:2021/12/31[pdat]))"""


dict_for_paper = {"pmid": [], "url": [], "journal": [], "title": [], "doi": [], "year": [], "majorMESH": [],
                  "otherMESH": [], "Authors": [], "filliation": [], "article_type": [], "abstract": []}
fetch = PubMedFetcher()
nlist = fetch.pmids_for_query(query, retmax=50000)  # default gets 250
for pmid in nlist:


    article = fetch.article_by_pmid(pmid)
    # print(article)
    # print(props(article))
    # print(article.authors)
    doi = article.doi
    title = article.title
    year = article.year
```

```
url = article.url
authors = article.authors_str
journal = article.journal
abstract = article.abstract
dict_for_paper["doi"].append(doi)
dict_for_paper["Authors"].append(authors)
dict_for_paper["year"].append(year)
dict_for_paper["title"].append(title)
dict_for_paper["url"].append(url)
dict_for_paper["pmid"].append(pmid)
dict_for_paper["journal"].append(journal)
dict_for_paper["abstract"].append(abstract)

#     print(article.author1_last_fm)
#    print(article.author1_lastfm)
# print(article.mesh)
other_l = []
mjrMesh = ""
for meshtermid, descr in article.mesh.items():
    #     print( meshtermid,descr )
    if descr["major_topic"]:
        mjrMesh = meshtermid
    else:
        other_l.append(meshtermid)

dict_for_paper["majorMESH"].append(mjrMesh)
dict_for_paper["otherMESH"].append(";".join(other_l))

filliation_l = []
```

```
    root = ET.fromstring(article.xml)
    try:
        author_list = root[0][0][3][5]  # AuthorList
    except Exception:
        pass
    for author in author_list:
        try:
            # print(author)
            filliation = author.find('AffiliationInfo')
            if filliation is not None:
                for f in filliation:
                    filliation_l.append(f.text)
        except Exception:
            pass

    dict_for_paper["filliation"].append(";".join(filliation_l))

    article_type = "|".join(article.publication_types.values())
    dict_for_paper["article_type"].append(article_type)

try:
    df = pd.DataFrame.from_dict(dict_for_paper)
except:
    print(dict_for_paper)

# print(df)
df.to_csv("test16.csv", index=False)
```