Evaluating the DFCM Research Program

A Computational Bibliometric Analysis

Christopher Meaney

Biostatistician, Department of Family and Community Medicine, UofT

March 12, 2019

Acknowledgements

- Dr. Eva Grunfeld for championing work and presenting at DFCM Core Exec Meeting.
- Dr. Paul Krueger for providing DFCM AAS publication data for Figure 1.
- Dr. David White for helping to disambiguate DFCM author names/IDs.
- Ms. Holly Downey and Ms. Julia Baxter for validating Scopus IDs.
- Ms. Caroline Turenko and Ms. Sarah Letovsky for helpful discussions on FOCUS DB.
- Drs. R. Moineddin, R. Freeman, M. Kulasegaram for helpful comments.

Evaluating the DFCM Research Program

- **Question:** How do stakeholders evaluate research contributions of DFCM faculty members?
- **Current Approach:** Annually survey research contributions from DFCM faculty members.
- **An Alternative Approach:** Bibliometric methods to extract knowledge DFCM research pgm.

A Selection of Questions we Can Investigate with Bibliometric Methods

- **Analysis of Citation Data**
  - (Historically), what is the total number of publications by DFCM faculty?
  - How does the number of research publications by DFCM faculty change over time?
  - Can we describe the origins of research at DFCM (i.e. earliest publications)?
  - What academic journals do DFCM faculty select for publishing their research?
  - Which faculty members at DFCM are publishing more/less research?
  - Does there exist variation in research productivity across DFCM hospital sites?
  - Can we describe the research productivity of DFCM-funded researchers?
- **Analysis of Abstract Data**
  - What language do DFCM researchers use in abstracts (unigram/bigram statistics)?
  - What emergent (probabilistic) topics best describe the DFCM abstract corpus?
  - Can unsupervised machine learning help to organize/browse DFCM abstracts?
- **Analysis of Authorship Data**
  - Who are the most frequent collaborators at DFCM?
  - Which non-DFCM faculty are we most likely to collaborate with?
  - Can recommender systems predict fruitful collaborative opportunities?

A General Computational Bibliometric Methodology

- We use direct database API-calls and/or web scraping to obtain publication data.
- We interact with these databases using explicit syntactic queries.
- The result of a query is an array of pub data (all publication data for faculty member).
- The aggregation of faculty-member-level data objects yields DFCM object.
- Curation of citation/abstract/authorship data permits description/modelling.
- Quantitative description and/or modelling permit knowledge discovery.

Scholarly Databases for Bibliometric Research

- **PubMed/NLM:** Query as `"Author First + Author Last[AU]"`. (Analysis Complete).
- **Scopus:** Query by persistent author ID. (Analysis Complete).
- **Web of Science:** Query by persistent author ID. (Analysis Not High Value).
- **Google Scholar:** Query by persistent author ID. (Analysis Blocked by Google).

Integrating FOCUS Data and Scholarly DB Data for Bibliometric Analyses

- FOCUS database most importantly defines population of DFCM faculty.
- FOCUS database provides info on start/end dates as incl/excl criteria.
- FOCUS also provides demographic information on faculty members.

Querying Publication Databases

Scopus Database: Author ID Search Queries

- N=2,527 unique faculty (plus staff, students, etc.) identified in FOCUS.
- N=1,195 DFCM faculty have first/last name associated with Scopus ID.
- N=521 DFCM faculty with likely Scopus ID (CM name disambiguation).
- N=220 DFCM faculty confirmed by DGW and CM as part DFCM.
- N=220 DFCM faculty produce N=4,315 publications between DFCM start/end date.

PubMed Database: Author Search String Queries

- Could in theory construct valid search string for N=2,517 faculty. Issue=Over-count.
- Use same subset N=220 faculty defined in Scopus analysis.
- N=184 (of 220) publish in PubMed, resulting in N=2,613 publications.
- Original query over-counts: `"Author First + Author Last[AU]"`.
- Revised query conservative: `"Author First + Author Last[AU] and Toronto[AD]"`.

Web of Science Database: Author ID Search Queries

- Issue: Faculty members must sign up for Web of Science ID. Few have ID.
- Using subset DFCM faculty who have Web of Science ID would under-count.

Google Scholar Web Scraping: Author ID Search Queries

- Issue: Faculty members must sign up for Google Scholar ID. Few have ID.
- Issue: Google sets throttle limits on IP calls to server. Blocked IP.

Statistical Methods

Analysis of Citation Data

- Simple counts/percentages described many different aspect DFCM research program.

Analysis of Abstract Data

- Transform text data (i.e DFCM abstracts) into numeric array for statistical analyses.
- Mathematical structures: document term matrices and term co-occurrence matrices.
- Simple descriptive statistics on NLP matrices provide insights language.
- Latent Dirichlet Allocation, Matrix Factorization, etc. for topic modelling.

Analysis of Authorship Data

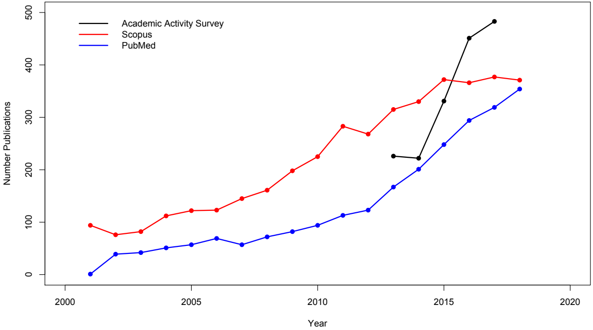
- Transform co-authorship data into a network adjacency matrix/graph.
- Descriptive methods applied adjacency matrix imply network characteristics.
- Matrix factorization for recommender engines (TBD).
- Probabilistic stochastic block models for community detection (TBD).

Estimating the Total Number of Research Publications by DFCM Faculty

- Scopus: **N=220** researchers produce **4,315** unique publications while at DFCM.
- PubMed: **N=184** researchers produce **2,613** unique publications while at DFCM.

Trends in the Number of DFCM Research Publications Over Time

- All methods suggest an increase in the number of publications over time.
- Scopus estimates greater number publication than PubMed across years.
- Both Scopus/PubMed track AAS publications estimates (relatively) closely.



The Origins of Primary Care Research at DFCM

- Early publications date back to 1984 (35 years ago).
- Exploration reveals early DFCM researchers have diverse interests:
  - Education scholarship.
  - Faculty development.
  - Global Health.
  - The practice of family medicine.
  - Obstetrics in family medicine practice.
- Some original research. Some expert commentary.

A Selection of Early Publications by DFCM Researchers

Primary Author	Year	Journal Title	Article Title
Rubenstein, W.	1984	Acad Med	Nutrition Ed. in FP Residency.
Talbot, Y.	1984	Acad Med	Fac Dev in FM: Survey and Needs Assessment.
Ovens, H.	1986	CMAJ	Anapylaxis due to vaccination in the office.
Fallis, G.	1988	J Rural Hlth	Small Hospital Obstetrics: Is Small Beautiful?
Berger, P.	1988	CMAJ	Offensive HIV legislation in Ontario.
Ellison, P.	1988	CMAJ	Endometrial carcinoma: Case-finding in FP's office
Fallis, G.	1989	J Fam Prac	Obstetric outcomes in rural practice.
Fallis, G.	1989	J Pub Hlth	Comparison obirthweights rural Zaire and Ontario.
Rubenstein, W.	1989	NEJM	Resource-Based Scale for Physicians' Reimbursement.
Carroll, J.	1990	ObGyn Survey	Differences obstetric care by FP and ObGyn.
White, D.	1991	CMAJ	Wearing wife-assault-prevention button: Impact on FPs.
Biringer, A.	1994	CFP	Psychosocial risk factors during pregnancy.

Where do DFCM Faculty Publish their Research?

- Canadian family medicine, and general medicine journals most common (i.e. CFP, CMAJ).
- National medical specialty journals (e.g. J. ObGyn Can, CJEM, Can J Cardio, etc.).
- Medical education journals (e.g. Med Ed, Acad Med, etc.).
- Open access journals (e.g. BMC family, PLoS family, BMJ Open, CMAJ Open, etc.).
- High impact (international) medical journals (e.g. NEJM, BMJ, Lancet, etc.).

Top-10 Journals by Number of Publications for DFCM Researchers

Rank	Scopus	N=4,315	PubMed	N=2,613
1	CFP	N=640	CFP	N=308
2	CMAJ	N=236	CMAJ	N=112
3	J ObGyn Can	N=111	PLoS ONE	N=62
4	PLoS ONE	N=80	J ObGyn Can	N=60
5	CJEM	N=60	BMJ Open	N=41
6	Can J Pub Hlth	N=44	CMAJ Open	N=40
7	BMJ Open	N=41	Acad Med	N=39
8	Can J Cardio	N=41	CJEM	N=36
9	Imp Sci	N=39	Med Ed	N=31
10	Med Ed	N=39	Can j Cardio	N=31

Journal titles suggest DFCM research has national/international impact.

Who is Generating Greatest Number of Research Publications at DFCM?

- Measure of publication volume using simple counts.
- Does not consider article type, article impact, downstream citations, etc.
- Does not account/adjust for person-years in organization.
- Somewhat biased against new faculty (only counts DFCM contributions).

Rank	Scopus	N=4,315	PubMed	N=2,613
1	Moineddin, R.	N=309	Moineddin, R.	N=272
2	Upshur, R.	N=288	Glazier, R.	N=204
3	Glazier, R.	N=275	Upshur, R.	N=186
4	Lexchin, J.	N=236	Le Foll, B.	N=127
5	Carroll, J.	N=161	Selby, P.	N=117
6	Kahan, M.	N=160	Lexchin, J.	N=113
7	Le Foll, B	N=159	Ivers, N.	N=112
8	Tu, K.	N=135	Tu, K.	N=105
9	Ivers, N.	N=124	Carroll, J.	N=61
10	Kwong, J.	N=122	Grunfeld, E.	N=59

Variation in Research Volume Across Hospital Sites

- Simple count of number of publications by faculty currently at given hospital site.
- Academic teaching sites produce majority of research.
- Opportunity exists for intervention (research mentorship) in community sites.
- Issue: Counts based on faculty members **current** affiliation. FPs change affiliations.
  - Example 1: Do you count Lynn Wilson's publications as SJHC or WCH?
  - Example 2: Do you count Onil Bhattacharyya publications as SMH or WCH?

Rank	Scopus	N=4,315
1	TWH	N=742
2	WCH	N=686
3	SMH	N=595
4	MSH	N=595
5	Bridgepoint	N=274
6	CAMH	N=257
7	SHSC	N=189
8	NYGH	N=188
9	TEGH	N=74
10	SJHC	N=57
11	TSH	N=45
12	Trillium	N=28
13	MKSH	N=24
14	Baycrest	N=6
15	Southlake	N=5

Characterizing Publication Counts of DFCM-Funded Research Faculty.

- Research program has "funded" (at least) N=30 faculty for research.
- These faculty have produced **N=1533/4315 (36%)** total publications.
- Issue: pubs not necessarily generated during funded period, but while at DFCM.

Antoniou, T. (N=53)	Bhattacharyya, O. (N=68)	Borgundvaag, B. (N=61)
Butt, D. (N=32)	Carroll, J. (N=161)	Del Giudice, L. (N=3)
Dunn, S. (N=74)	Finkelstein, M. (N=80)	Glazier, R. (N=275)
Greiver, M. (N=79)	Heisey, R. (N=23)	Ivers, N. (N=124)
Jaakkimainen, L. (N=38)	Kahan, M. (N=160)	Kiran, T. (N=31)
Kwong, J. (N=122)	Landes, M. (N=23)	Lofters, M. (N=51)
McIsaac, W. (N=62)	Oandasan, I. (N=56)	Persaud, N. (N=29)
Pinto, A. (N=15)	Selby, P. (N=110)	Sodhi, S. (N=21)
Spithoff, S. (N=13)	Srivastava, A. (N=26)	Steele, L. (N=33)
Tu, K. (N=135)	Varner, C. (N=11)	Wentlandt, K. (N=17)

DFCM Abstract Corpus

- N=4,315 unique Scopus publications. Only 2,994 (69%) publications have abstract.
- Total of 385,872 words in the corpus (after removal stop words, low freq words, etc.).
- Number of unique words in corpus is 7,965 (single word = unigram).
- Number of unique bigrams (consecutive word pairs) is 63,066.

Most-Frequent Words/Unigrams and Bigrams in DFCM Abstract Corpus

- Simple unigram/bigram statistics across corpus give sense what publications are about.
- Strong face validity. Observed word frequency statistics as expected for DFCM corpus.

Rank	Word/Unigram	Frequency	Bigram	Frequency
1	care	N=4221	primary care	N=925
2	patients	N=4219	health care	N=771
3	health	N=3749	family physicians	N=570
4	study	N=2798	rights reserved	N=368
5	data	N=1853	public health	N=368
6	women	N=1808	family medicine	N=332
7	methods	N=1796	confidence interval	N=324
8	risk	N=1773	breast cancer	N=268
9	ci	N=1627	mental health	N=259
10	physicians	N=1590	outcome measures	N=239

DFCM Abstract Corpus

- N=2,994 abstracts. N=7,965 unique words (N=385,872 total words in corpus).
- Probabilistic topic models extend MF to learn latent thematic structure document corpora.
- Each document is mixture of topics. Each topic is categorical distribution over vocabulary.
- Loading of words on V-dimensional topical vectors used to describe corpus.
- K-dimensional topical mixture weights can be used for browsing/organization documents.

Top-7 Words Loading on 14-Selected Topical Vectors from DFCM Corpus

- Topics about research methods: design, epi, biostat, QL, KT/KTE, reviews/synthesis, etc.
- Topics about diseases: mental health, women's health, cardiology, diabetes, vaccines, etc.
- We fit K=100 topic LDA model to DFCM abstract corpus, can explore additional domains.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7
1	studies	articles	review	search	medline	embase	cochrane
2	intervention	trial	group	randomized	control	cluster	arm
3	interviews	qualitative	themes	focus	barriers	semi-structured	roles
4	medicine	residents	faculty	teaching	training	curriculum	trainee
5	cancer	screening	breast	colorectal	cervical	survivors	mammogram
6	prevalance	persons	population	rates	incidence	trends	mortality
7	women	pregnant	pre-natal	postpartum	antenatal	maternity	trimester
8	pain	opioid	methadone	neuropathic	cannabinoids	chronic	relief
9	survival	ems	arrest	resuscitation	cardiac	ohca	cpr
10	factors	socioeconomic	poor	ses	deprivation	health	neighbourhood
11	pressure	blood	levels	ses	deprivation	health	neighbourhood
12	drugs	pharmaceutical	industry	safety	medicines	products	regulatory
13	air	pollution	neighbourhood	no2	walkability	census	spatial
14	mental	disorder	dementia	psychiatric	illness	anxiety	emotional

Co-Authorship Data

- N=220 DFCM authors generate 4,315 publications.
- N=13,263 unique authors on publications. Implies 13,043 unique non-DFCM collaborators.
- Strong face validity to internal-DFCM collaborations.
- Frequent collaborations with non-DFCM experts in: HSR, AIDS, KTE, Cancer, Vaccine, etc.

Most-Frequent Research Collaborators within DFCM

Rank	Researcher 1	Researcher 2	Collaboration Frequency
1	Moineddin, R.	Glazier, R.	N=57
2	Moineddin, R.	Upshur, R.	N=31
3	Srivastava, A.	Kahan, M.	N=24
4	Antoniou, T.	Burchell, A.	N=21
5	Tu, K.	Ivers, N.	N=21
6	Tu, K.	Butt, D.	N=20
7	Grunfeld, E.	O'Brien, M.	N=18
8	Glazier, R.	Steele, L.	N=17
9	Wilson, L.	Kahan, M.	N=16
10	Glazier, R.	Lofters, A.	N=16

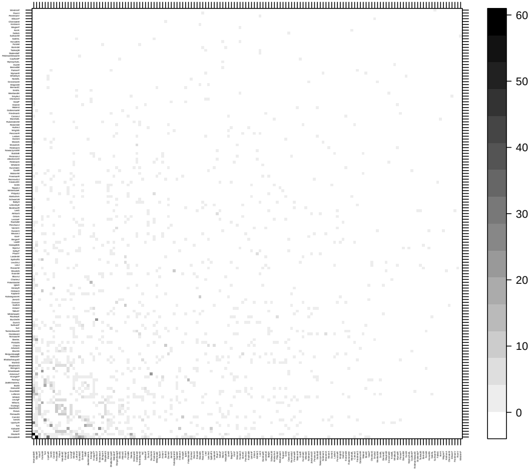
Whom From Outside DFCM Do We Collaborate with Most Often?

Mamdani, M (N=83)	Raboud, J. (N=73)	Grimshaw, J. (N=61)	Rachlis, A. (N=56)	Strauss, S. (N=49)
Zwarenstein, M. (N=48)	Bayoumi, A. (N=45)	Morrison, L. (N=44)	Crowcroft, N. (N=40)	McGeer, A. (N=36)

Co-Authorship Data

- Subset N=150 DFCM faculty members who co-author with other DFCM faculty members.
- Dimension adj matrix: (150,150). N=22,500 elements. N=1312 non-zero. 94% sparsity.
- A total of N=3314 collaborative authorships between DFCM faculty members.

Co−Authorship Network Adjacency Matrix



A Computational Bibliometric Review of the DFCM Research Program

- Analyzing citation/abstract/co-authorship data provides insights DFCM research program.
- Methodology generally feasible. Structured data enables answering multitude queries.
- Limitations Methodology:
  - Disambiguation of names/IDs to operationalize "DFCM" challenging.
  - Only get publication info from aggregator indexed journals. (Good idea IMO).
  - Cost to obtain bibliometric IDs. Cost to maintain DB of IDS in FOCUS.
  - Cost to update and run scripts. APIs/websites change, software brittle.
  - GS blocks IP. Scopus/WoS avail through API key. PubMed avail, but worst.
- Compare/Contrast with Current Approach (AAS):
  - AAS is valid/reliable survey; strong methodology. Generally trust pub count stats.
  - AAS suffers survey non-response issues. Possible undercount of publication stats.
  - Authors self-report publications (pubs not indexed in any sense). Possible over-count.
  - Citation data standardized by humans. Data quality? Human-time costs?
  - AAS citation/authorship data difficult to parse/query. No abstract data from AAS.

Concluding Thoughts

- Bibliometric methodology feasible for DFCM research program evaluation/review.
- Bibliometric methods provide interesting perspectives on DFCM research program.
- Some risks to bibliometric approach: costs, data access, name/ID disambiguation.