Semi-Automating Knowledge Base Construction for Cancer Genetics

Somin Wadhwa Kanhua Yin Kevin S. Hughes Byron C. Wallace

Northeastern University
Khoury College of Computer Sciences

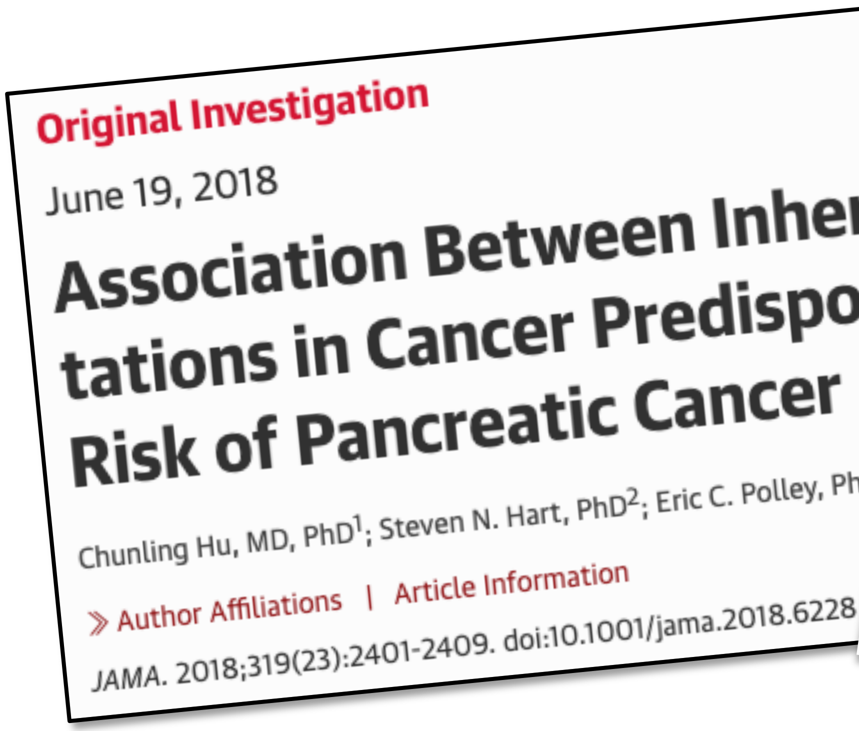
UMassAmherst
College of Information & Computer Sciences

MGH 1811 MASSACHUSETTS GENERAL HOSPITAL

HARVARD MEDICAL SCHOOL

Appraising oncology literature

Penetrance papers Medical literature describing risk of cancer with a particular pathogenic variant in cancer susceptibility gene.



Original Investigation

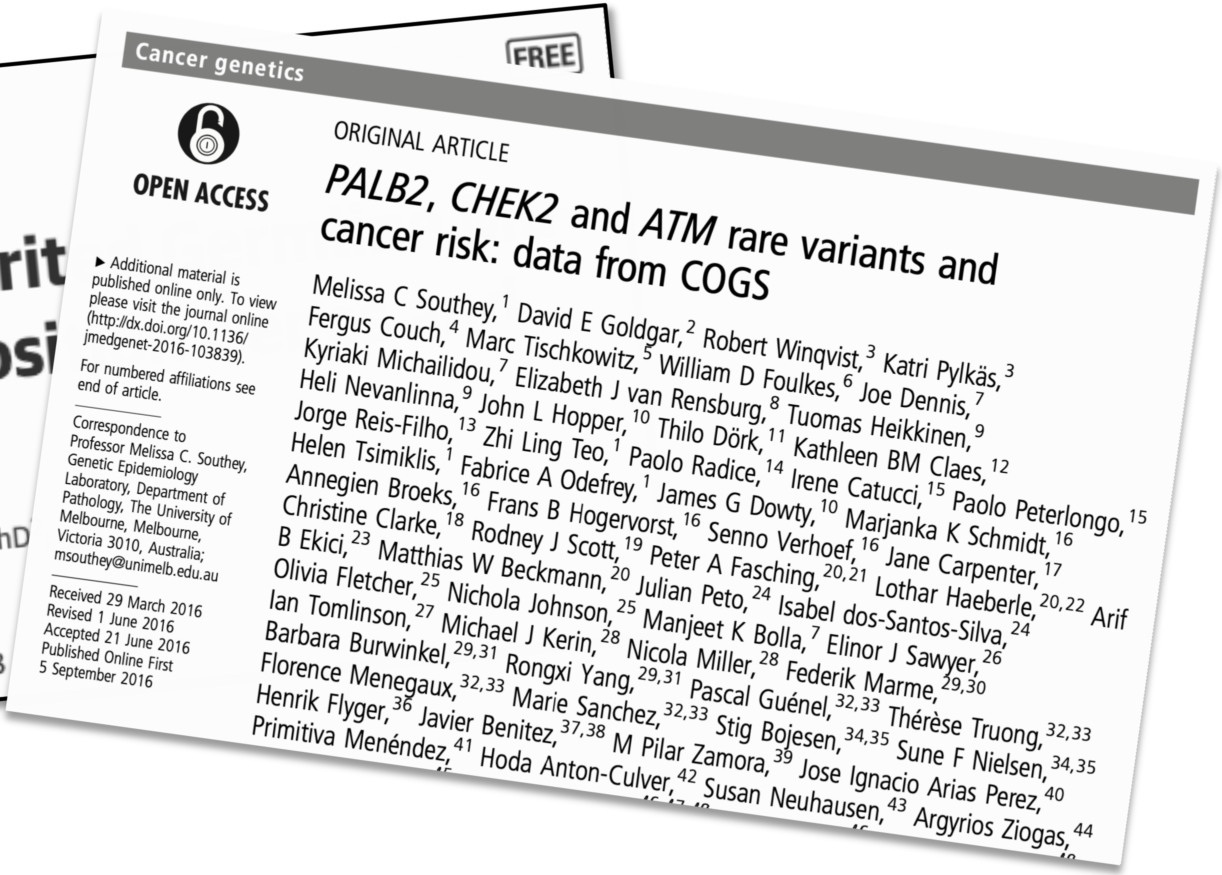
June 19, 2018

Association Between Inherit tations in Cancer Predisposi Risk of Pancreatic Cancer

Chunling Hu, MD, PhD[1]; Steven N. Hart, PhD[2]; Eric C. Polley, PhD

» Author Affiliations | Article Information

JAMA. 2018;319(23):2401-2409. doi:10.1001/jama.2018.6228



Cancer genetics FREE

OPEN ACCESS

ORIGINAL ARTICLE

PALB2, *CHEK2* and *ATM* rare variants and cancer risk: data from COGS

Melissa C Southey,[1] David E Goldgar,[2] Robert Winqvist,[3] Katri Pylkäs,[3] Fergus Couch,[4] Marc Tischkowitz,[5] William D Foulkes,[6] Joe Dennis,[7] Kyriaki Michailidou,[7] Elizabeth J van Rensburg,[8] Tuomas Heikkinen,[9] Heli Nevanlinna,[9] John L Hopper,[10] Thilo Dörk,[11] Kathleen BM Claes,[12] Jorge Reis-Filho,[13] Zhi Ling Teo,[1] Paolo Radice,[14] Irene Catucci,[15] Paolo Peterlongo,[15] Helen Tsimiklis,[1] Fabrice A Odefrey,[1] James G Dowty,[10] Marjanka K Schmidt,[16] Annegien Broeks,[16] Frans B Hogervorst,[16] Senno Verhoef,[16] Jane Carpenter,[17] Christine Clarke,[18] Rodney J Scott,[19] Peter A Fasching,[20,21] Lothar Haeberle,[20,22] Arif B Ekici,[23] Matthias W Beckmann,[20] Julian Peto,[24] Isabel dos-Santos-Silva,[24] Olivia Fletcher,[25] Nichola Johnson,[25] Manjeet K Bolla,[7] Elinor J Sawyer,[26] Ian Tomlinson,[27] Michael J Kerin,[28] Nicola Miller,[28] Federik Marme,[29,30] Barbara Burwinkel,[29,31] Rongxi Yang,[29,31] Pascal Guénel,[32,33] Thérèse Truong,[32,33] Florence Menegaux,[32,33] Marie Sanchez,[32,33] Stig Bojesen,[34,35] Sune F Nielsen,[34,35] Henrik Flyger,[36] Javier Benitez,[37,38] M Pilar Zamora,[39] Jose Ignacio Arias Perez,[40] Primitiva Menéndez,[41] Hoda Anton-Culver,[42] Susan Neuhausen,[43] Argyrios Ziogas,[44]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2016-103839).

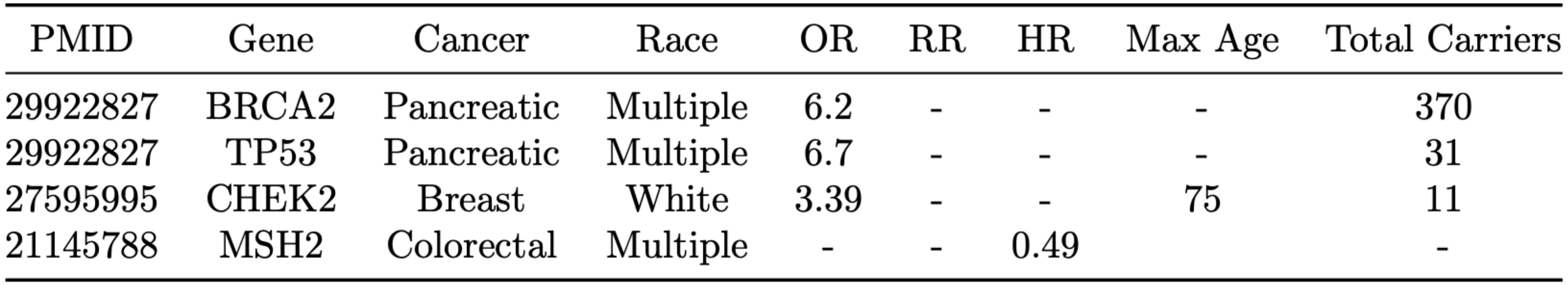
For numbered affiliations see end of article.

Correspondence to Professor Melissa C. Southey, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Melbourne, Victoria 3010, Australia; msouthey@unimelb.edu.au

Received 29 March 2016
Revised 1 June 2016
Accepted 21 June 2016
Published Online First
5 September 2016

Key study information

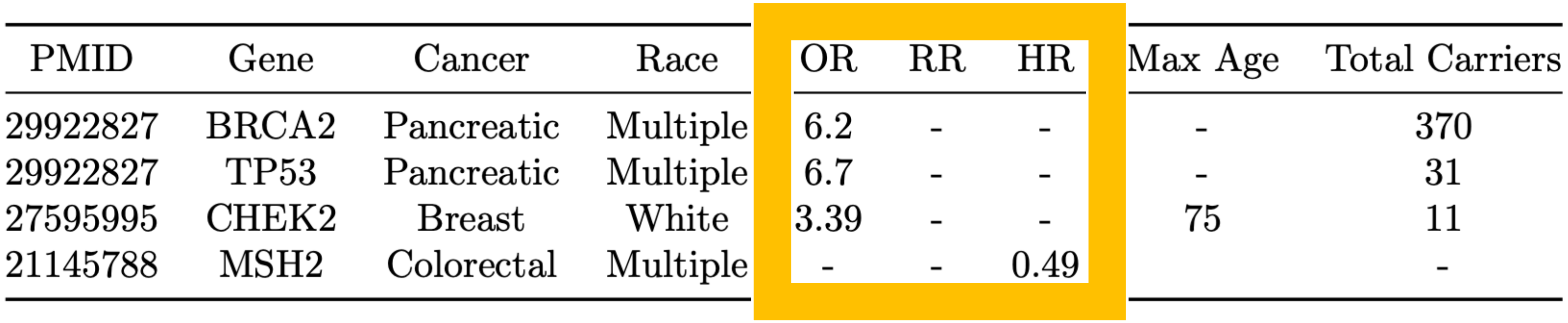
- What population was studied? ***Ascertainment***!
- How many patients were in the study?
- What cancer was the patient at-risk for? What was the associated risk?
- Ideally: Synthesize the key elements from papers into a database.
    - e.g. ask2me.org



PMID	Gene	Cancer	Race	OR	RR	HR	Max Age	Total Carriers
29922827	BRCA2	Pancreatic	Multiple	6.2	-	-	-	370
29922827	TP53	Pancreatic	Multiple	6.7	-	-	-	31
27595995	CHEK2	Breast	White	3.39	-	-	75	11
21145788	MSH2	Colorectal	Multiple	-	-	0.49		-

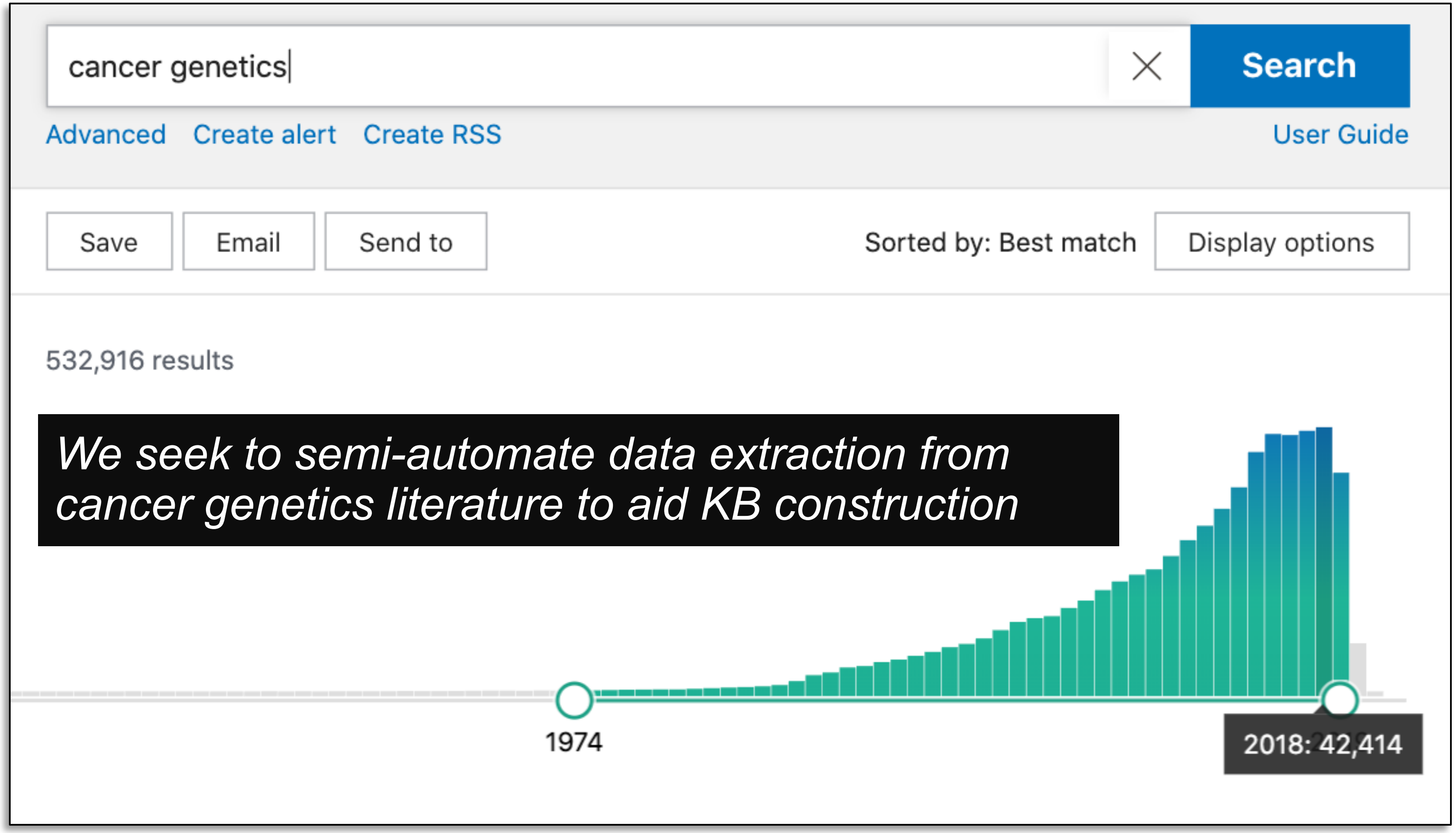
Key study information

- What population was studied? ***Ascertainment***!
- How many patients were in the study?
- What cancer was the patient at-risk for? What was the associated risk?
- Ideally: Synthesize the key elements from papers into a database.
    - e.g. ask2me.org



PMID	Gene	Cancer	Race	OR	RR	HR	Max Age	Total Carriers
29922827	BRCA2	Pancreatic	Multiple	6.2	-	-	-	370
29922827	TP53	Pancreatic	Multiple	6.7	-	-	-	31
27595995	CHEK2	Breast	White	3.39	-	-	75	11
21145788	MSH2	Colorectal	Multiple	-	-	0.49		-

The problem: Too many studies



cancer genetics

532,916 results

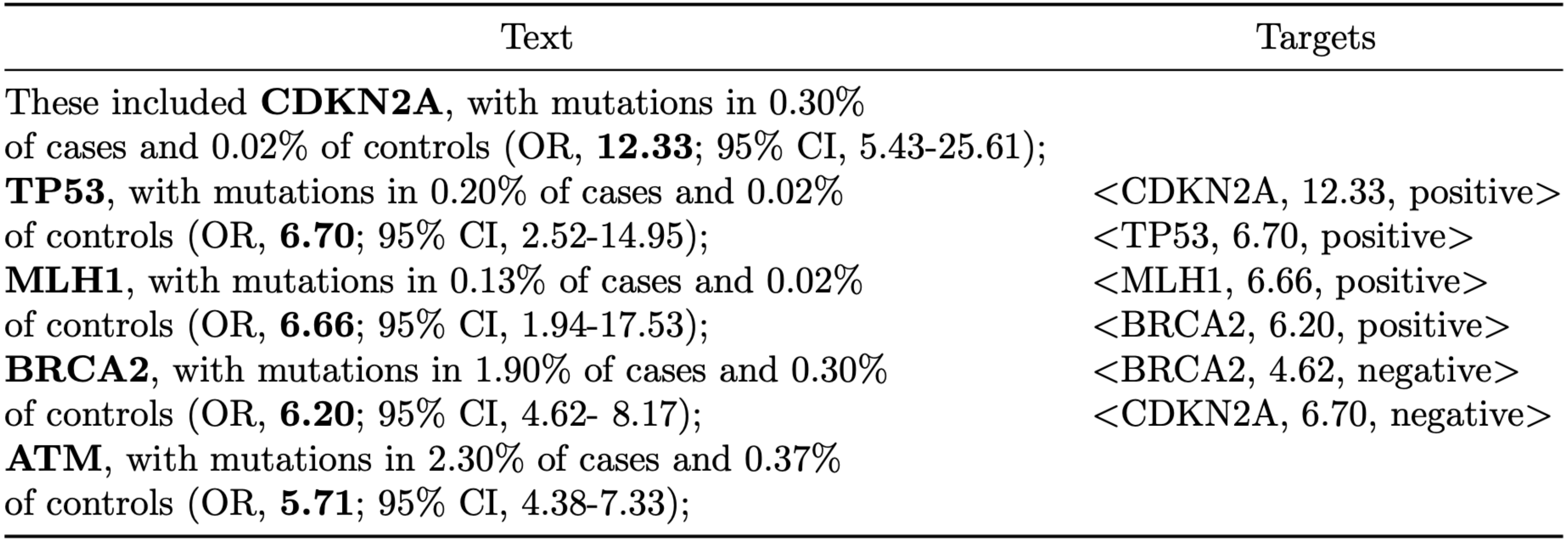
We seek to semi-automate data extraction from cancer genetics literature to aid KB construction

Ascertainment

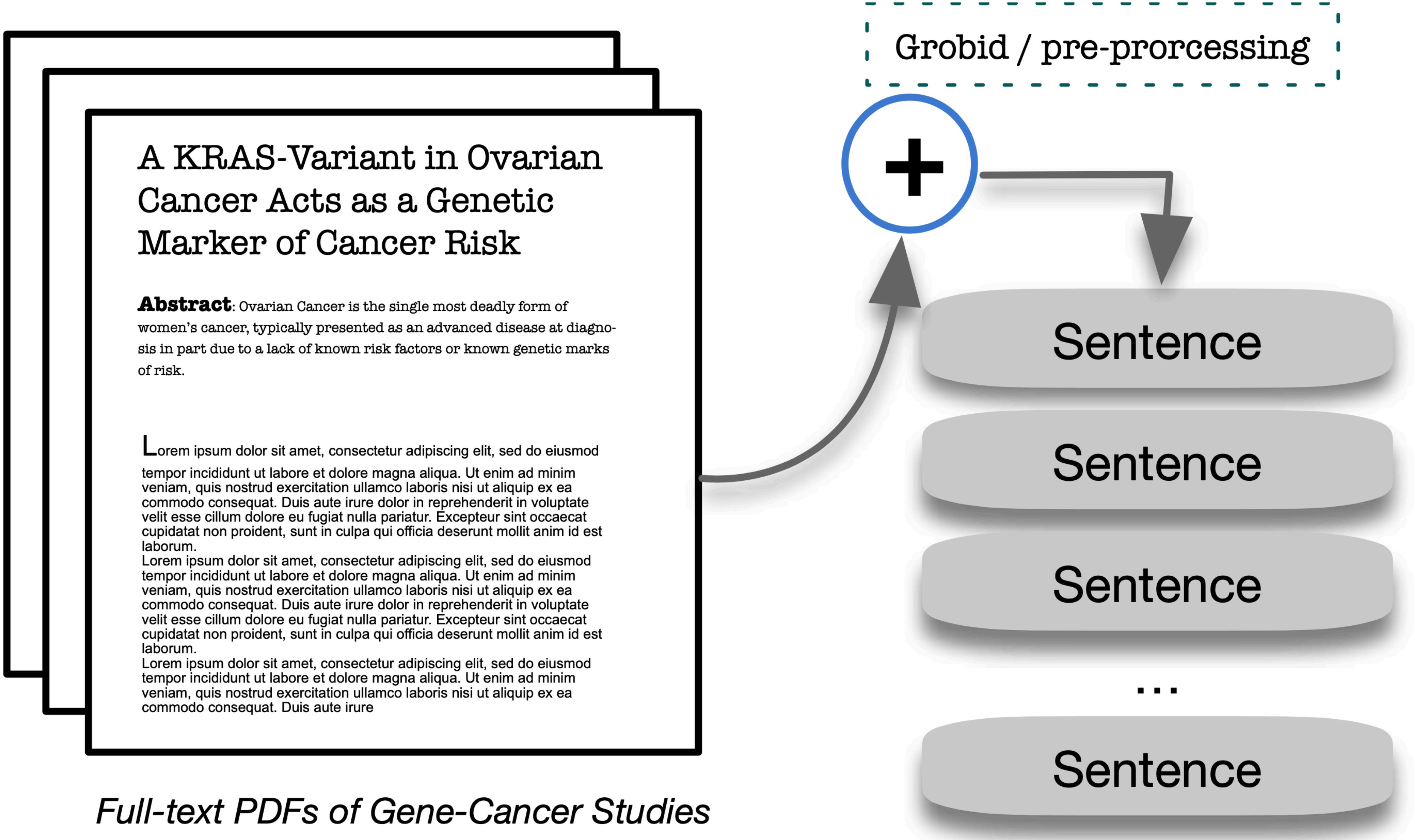
A control population was defined from the National Danish Civil Registration System, matched for sex, year of birth, mutation carriers as well as first degree relatives.

For age adjusted analysis, the projected U.S. population was used (year 2000); 84% of the 3499 individuals were white.

Risks



Text	Targets
These included **CDKN2A**, with mutations in 0.30% of cases and 0.02% of controls (OR, **12.33**; 95% CI, 5.43-25.61); **TP53**, with mutations in 0.20% of cases and 0.02% of controls (OR, **6.70**; 95% CI, 2.52-14.95); **MLH1**, with mutations in 0.13% of cases and 0.02% of controls (OR, **6.66**; 95% CI, 1.94-17.53); **BRCA2**, with mutations in 1.90% of cases and 0.30% of controls (OR, **6.20**; 95% CI, 4.62- 8.17); **ATM**, with mutations in 2.30% of cases and 0.37% of controls (OR, **5.71**; 95% CI, 4.38-7.33);	<CDKN2A, 12.33, positive> <TP53, 6.70, positive> <MLH1, 6.66, positive> <BRCA2, 6.20, positive> <BRCA2, 4.62, negative> <CDKN2A, 6.70, negative>



A KRAS-Variant in Ovarian Cancer Acts as a Genetic Marker of Cancer Risk

Abstract: Ovarian Cancer is the single most deadly form of women's cancer, typically presented as an advanced disease at diagnosis in part due to a lack of known risk factors or known genetic marks of risk.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure

Grobid / pre-prorcessing

Sentence

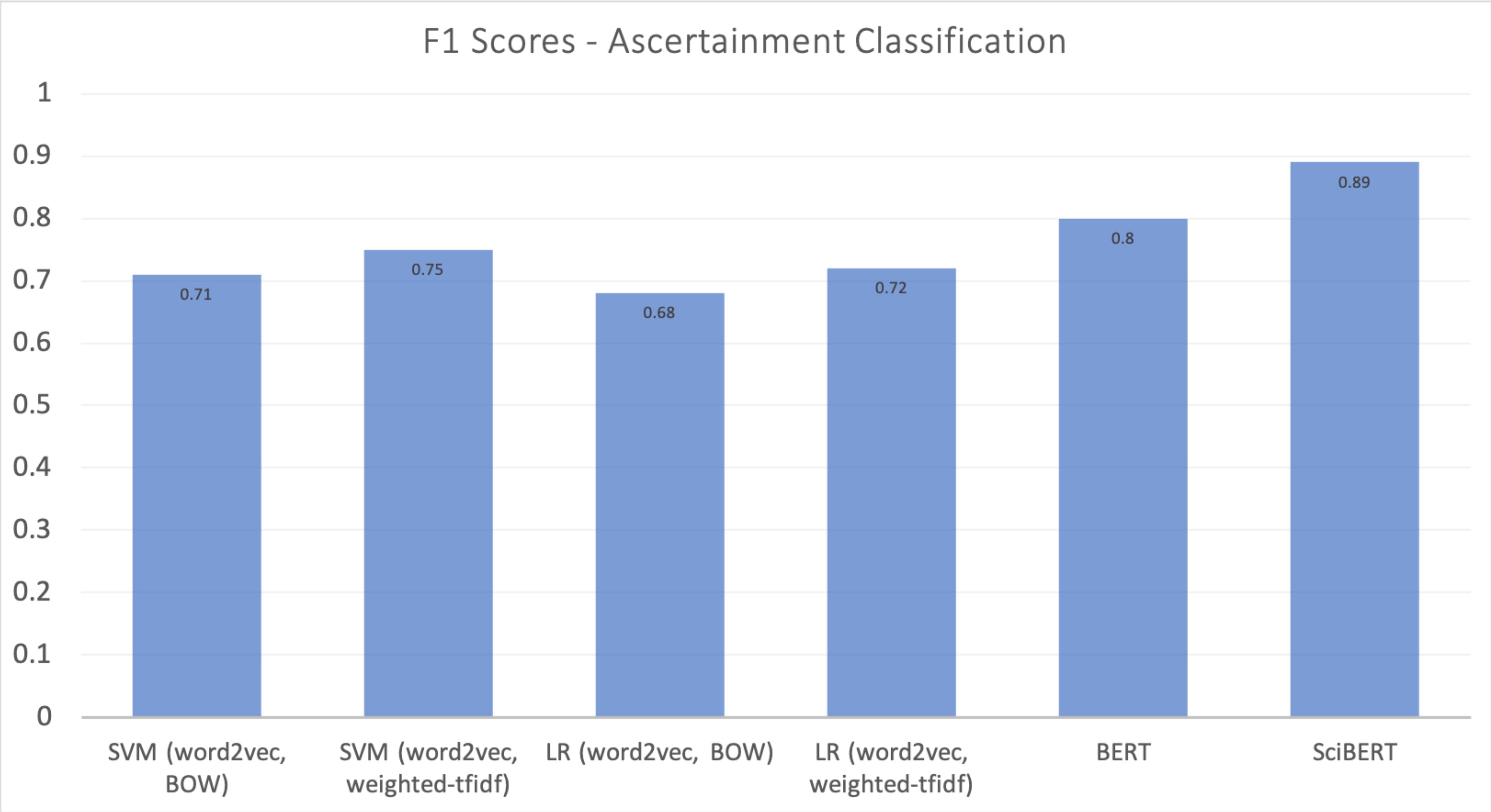
Sentence

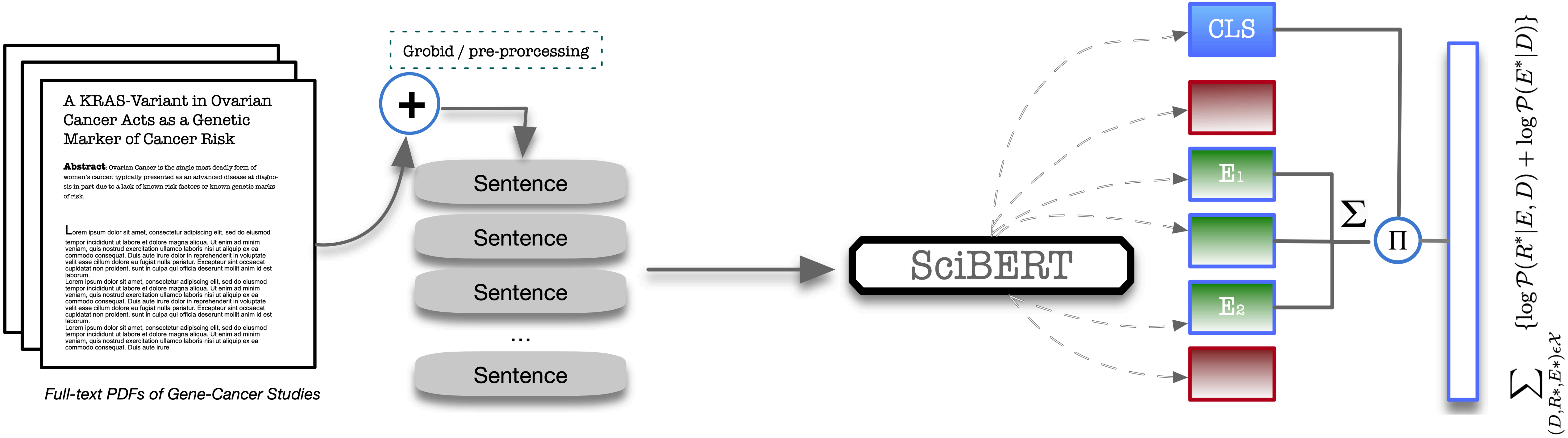
Sentence

...

Sentence

Full-text PDFs of Gene-Cancer Studies


F1 Scores - Ascertainment Classification
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.71
0.75
0.68
0.72
0.8
0.89
SVM (word2vec, BOW)
SVM (word2vec, weighted-tfidf)
LR (word2vec, BOW)
LR (word2vec, weighted-tfidf)
BERT
SciBERT



A KRAS-Variant in Ovarian Cancer Acts as a Genetic Marker of Cancer Risk

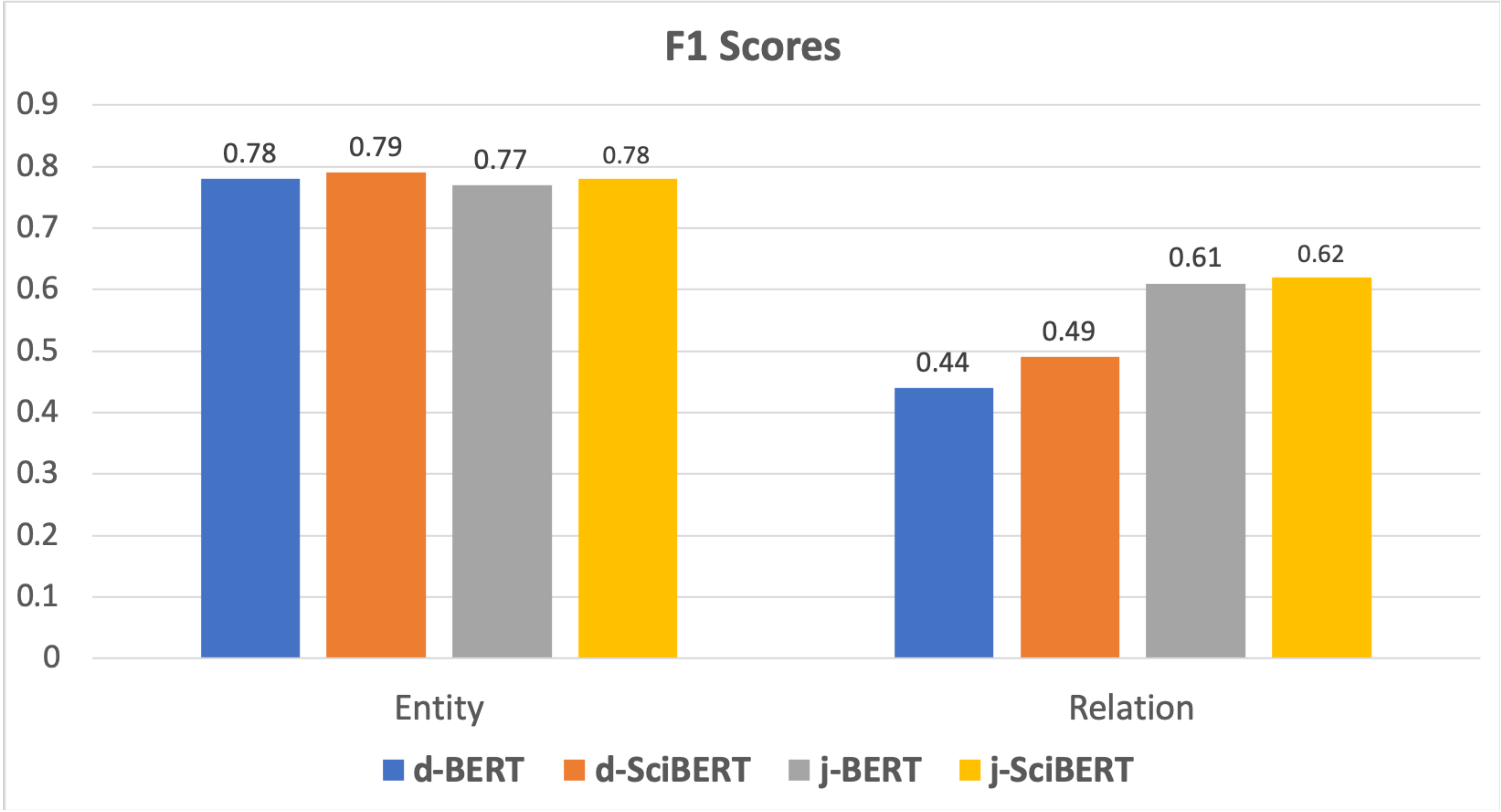
Abstract: Ovarian Cancer is the single most deadly form of women's cancer, typically presented as an advanced disease at diagnosis in part due to a lack of known risk factors or known genetic marks of risk.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure

Full-text PDFs of Gene-Cancer Studies

$$\sum_{(D,R*,E*)\epsilon\mathcal{X}} \{\log \mathcal{P}(R^*|E,D) + \log \mathcal{P}(E^*|D)\}$$

<germline-mutation, risk> <germline-mutation, risk> ⋯ <germline-mutation, risk>



d – disjoint; j – joint

Thank you!

sominwadhwa.com ||🐦@sominw || sominwadhwa@cs.umass.edu

Northeastern University
Khoury College of
Computer Sciences

UMassAmherst
College of Information
& Computer Sciences

MGH 1811
MASSACHUSETTS
GENERAL HOSPITAL

VE RI TAS
HARVARD
MEDICAL SCHOOL