The Geisinger DiscovEHR cohort: 233,185 people with exome sequencing and longitudinal EHR data

Douglas R. Stewart, MD

Jung Kim, Ph.D.

Clinical Genetics Branch

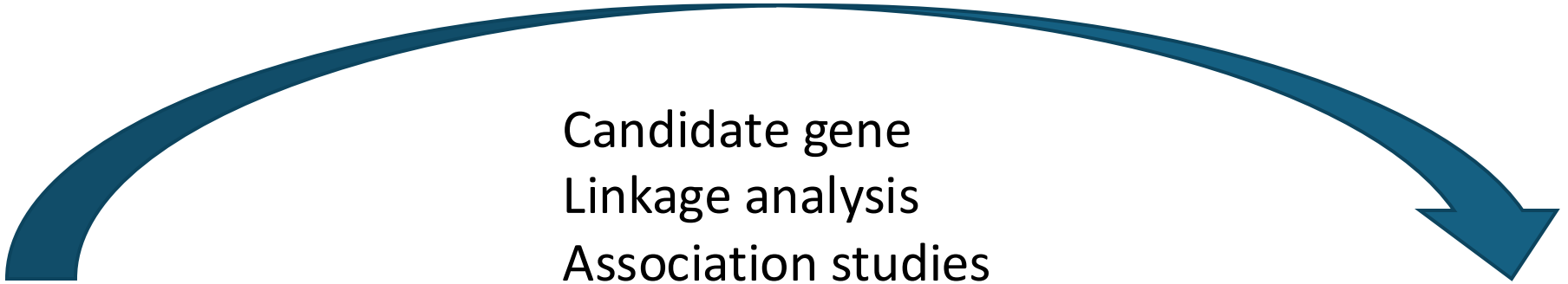
NIH NATIONAL CANCER INSTITUTE

January 8, 2025

Outline

1. *What is genomic ascertainment?*
2. *MyCode and DiscovEHR*
3. *CHEK2 cancer risk*
4. *RASopathies cancer risk*
5. *Lessons learned*
6. *Practical matters*

Phenotype-first model of clinical genetics…



Candidate gene
Linkage analysis
Association studies
Family structure

Phenotype
(Disease)

Individuals/Families
Smaller cohorts
Clinic-based
10^2 – 10^3 participants

Genome
(Germline variation)

Sanger
Panels
Exome/Genome
Arrays
Copy-number

Phenotype-First

"Ascertain weird phenotype *then* find genotype"

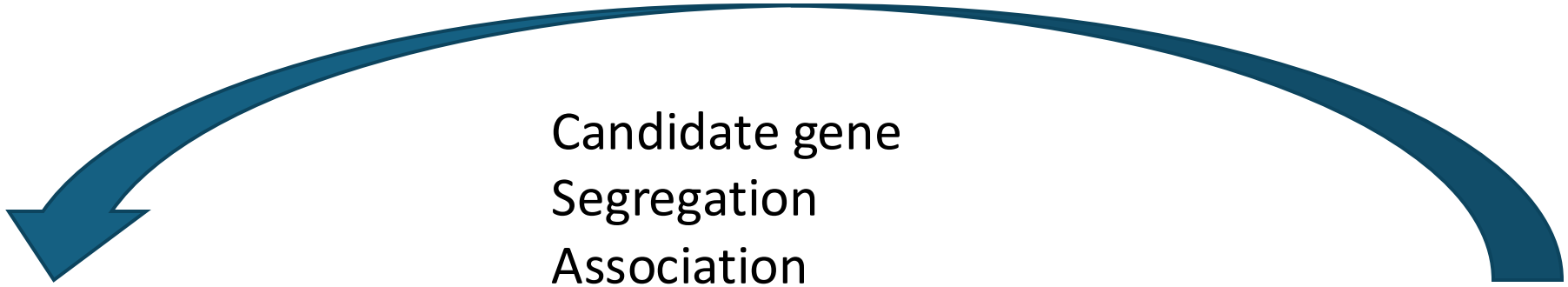
Strengths

- "Tried-and-true" approach to rare disease
- Builds on expertise of clinical investigator (pattern/syndrome recognition)
- Well-trod recruitment strategies
- Costs can be more modest (single clinic/investigator recruiting families)

Weaknesses

- Ascertainment biases
- Miss non-penetrant cases
- Miss rare or unknown manifestations of disorder
- May over-estimate severity of disorder
- Reactive
- Time- and labor-intensive to build special cohorts; lower throughput

Genome-first approach to clinical genetics…



Candidate gene
Segregation
Association
New methods?

Phenotype
(Disease)

Focused or broad
Health system-based
Electronic health record (EHR)
10^3 – 10^6 participants
Populations
Countries

Genome
(Germline variation)

Exome
Genome
Panels
(Population-scale cohorts)

Genotype (Genome)-First*

"Ascertain weird genotype *then* find phenotype"

Strengths

- See full phenotypic spectrum, especially at older ages
- Wider range of severity
- Better penetrance estimates
- Proactive
- Multiple gene/pathways
- Opportunities for syndrome discovery
- Higher throughput?

Weaknesses

- Different ascertainment biases
- $10^{-5} \times 10^{6} = 10^{1}$
- Infrastructure requirements
- Significant costs to build/recruit cohorts
- Bioinformatics expertise – variant classification
- Data science expertise
  - Clinical bioinformatics for phenotypes
  - Missing/sparse data (few clinical visits)
  - Quirks of medical coding
  - Medical coding: for billing, not research!

**AKA*: "Public health genomics," "Population genomics," "Reverse phenotyping," "Genomic ascertainment"

What are the consequences of genomic ascertainment?

- *Prevalence* of pathogenic/likely pathogenic (P/LP)* variants is (often) greater than previously estimated

- *Penetrance* (*risk* from a P/LP variant) may not be as high as previously estimated

- *Phenotype* is different (may be less severe, broader)

*Clinically actionable germline variation classified by ACMG/AMP rules (Richards *et al Genetics in Medicine* 2015)

Genetics inMedicine | ORIGINAL RESEARCH ARTICLE

and Genomics

Carey et al *Genet Med* 2016

The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research

David J. Carey, PhD[1], Samantha N. Fetterolf, BS[1], F. Daniel Davis, PhD[1], William A. Faucett, MS[1], H. Lester Kirchner, PhD[1], Uyenlinh Mirshahi, PhD[1], Michael F. Murray, MD[1], Diane T. Smelser, PhD[1], Glenn S. Gerhard, MD[2] and David H. Ledbetter, PhD[1]

Geisinger Medical Center, Danville, Pennsylvania
Opened 1915 as Geisinger Hospital

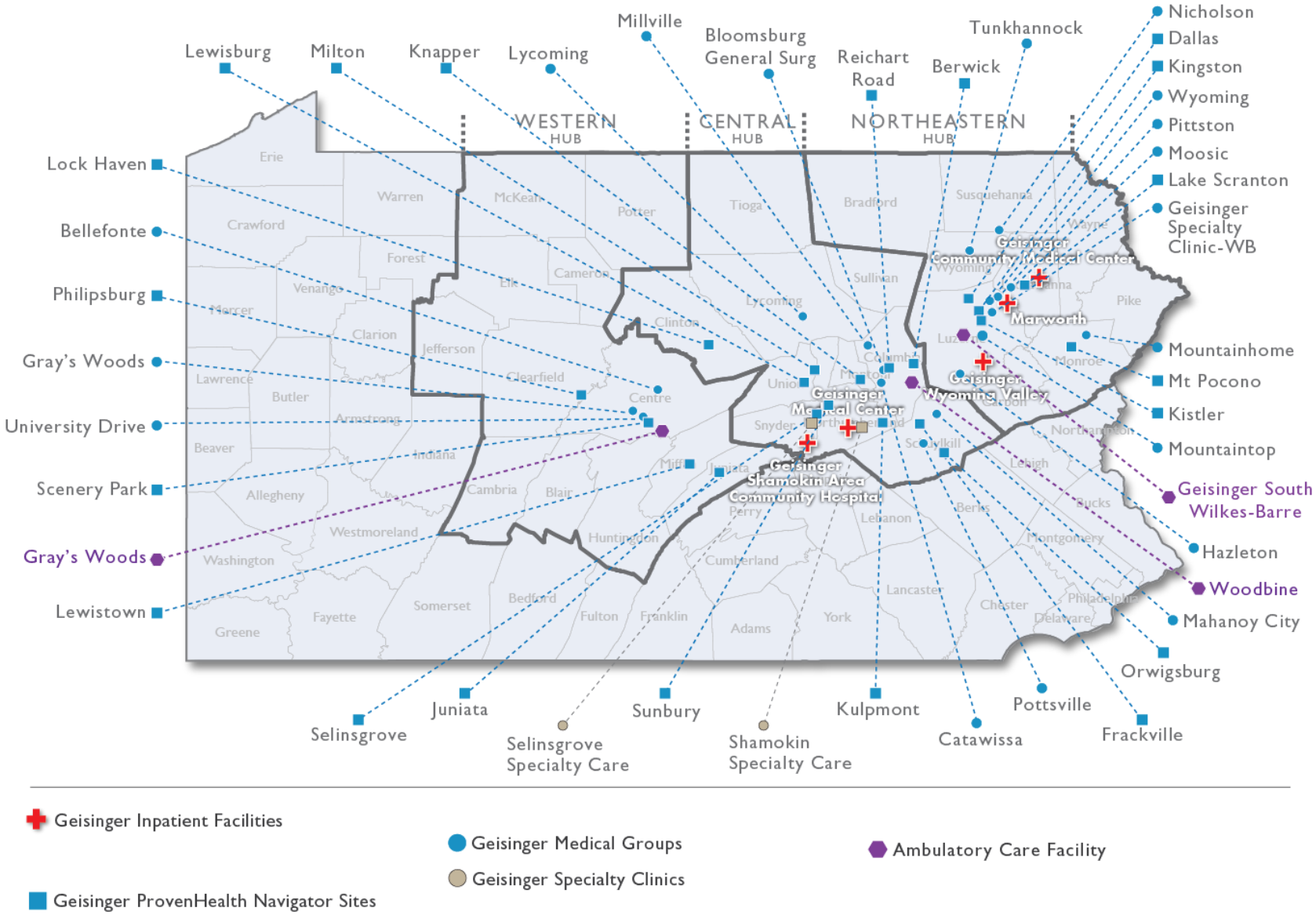
Geisinger

Abigail Geisinger
1827 - 1921

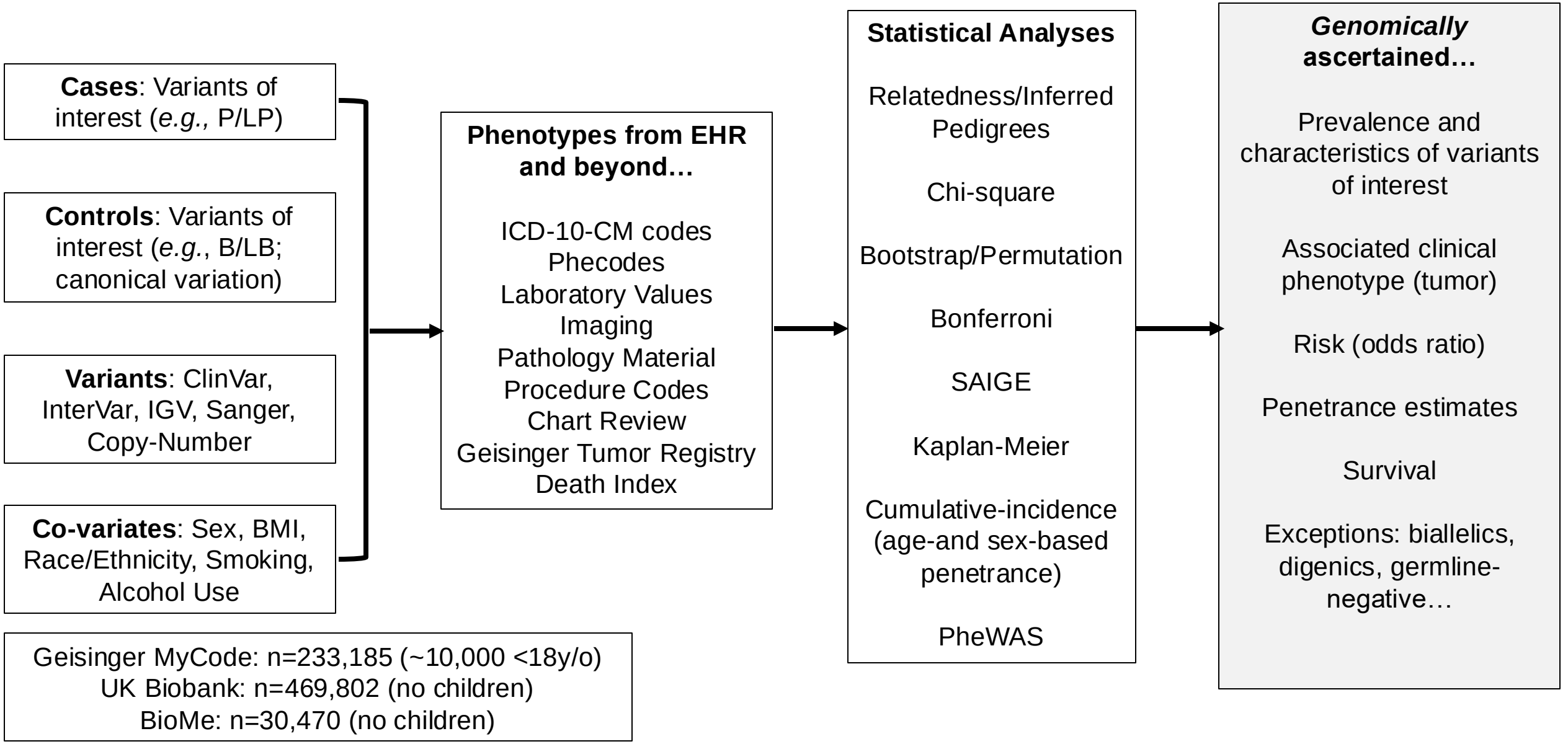
Genome-first approach using population-scale sequencing linked to electronic health records

Geisinger

- Serves >3M people
- Relatively non-transient; many multi-generation families; low in/out migration
- EHR since 1995
- 233,185 exomes (+/-arrays) (1/2025; with Regeneron)
- 5-year (2021-2026), DCEG-funded contract to investigate monogenic tumor predisposition disorders using the genome-first approach



Implementation of a genome-first approach in three population-sized, exome-sequenced, EHR-linked cohorts



Genetics in Medicine | ARTICLE

A genotype-first approach to exploring Mendelian cardiovascular traits with clear external manifestations

Brittany M. Wenger, BS[1], Nihir Patel, MS[2], Madeline Lui, BA[1], Arden Moscati, PhD[3], Ron Do, PhD[3], Douglas R. Stewart, MD[4], Marco Tartaglia, PhD[5], Laura Muiño-Mosquera, MD, PhD[6,7], Julie De Backer, MD, PhD[7,8], Amy R. Kontorovich, MD, PhD[2,9] and Bruce D. Gelb, MD[2,10]

HGG Advances | ARTICLE

Genome-first approach of the prevalence and cancer phenotypes of pathogenic or likely pathogenic germline *TP53* variants

Kelvin C. de Andrade,[1,7,*] Natasha T. Strande,[2] Jung Kim,[1] Jeremy S. Haley,[2] Jessica N. Hatton,[1] Megan N. Frone,[1] Payal P. Khincha,[1] Gretchen M. Thone,[2] Uyenlinh L. Mirshahi,[2] Cynthia Schneider,[3] Heena Desai,[3] James T. Dove,[2] Diane T. Smelser,[2] Penn Medicine BioBank,[6] Regeneron Genetics Center,[6] Arnold J. Levine,[4] Kara N. Maxwell,[3] Douglas R. Stewart,[1,5] David J. Carey,[2,5] and Sharon A. Savage[1,5]

Genetics in Medicine (2024) 26, 101042

ELSEVIER

Genetics in Medicine
An Official Journal of the ACMG
www.journals.elsevier.com/genetics-in-medicine

JAMA Network Open

Genetics in Medicine Open (2024) 2, 10

Research

ELSEVIER

JAMA Dermatology | Brief Report

Estimated Prevalence, Tumor Spectrum, and Neurofibromatosis Type 1–Like Phenotype of *CDKN2A*-Related Melanoma-Astrocytoma Syndrome

Michael R. Sargen, MD; Jung Kim, PhD; Thomas P. Potjer, MD, PhD; Mary E. Velthuizen; Arelis E. Martir-Negron, MD; Yazmin Odia, MD; Hildur Helgadottir, MD, PhD; Jessica N. Hatton, MS, CGC; Jeremy S. Haley, MS; Gretchen Thone, MS, CGC; Brigitte C. Widemann, MD; Andrea M. Gross, MD; Marielle E. Yohe, MD, PhD; Rosandra N. Kaplan, MD; Jack F. Shern, MD; R. Taylor Sundby, MD; Esteban Astiazaran-Symonds, MD; Xiaohong R. Yang, PhD, MPH; David J. Carey, PhD; Margaret A. Tucker, MD; Douglas R. Stewart, MD; Alisa M. Goldstein, PhD

BRIEF REPORT

Most Fanconi anemia heterozygotes are not at increased cancer risk: A genome-first DiscovEHR cohort population study

Joseph Deng[1], Burak Altintas[1,2], Jeremy S. Haley[3], Jung Kim[1], Mark Ramos[4], David J. Carey[3], Douglas R. Stewart[1], Lisa J. McReynolds[1,*]

ARTICLE

A genome-first approach to characterize *DICER1* pathogenic variant prevalence, penetrance and cancer, thyroid, and other phenotypes in 2 population-scale cohorts

Jung Kim[1], Jeremy Haley[2], Jessica N. Hatton[1], Uyenlinh L. Mirshahi[2], H. Shanker Rao[2], Mark F. Ramos[1], Diane Smelser[2], Gretchen M. Urban[2], Kris Ann P. Schultz[3,4,5], David J. Carey[2], Douglas R. Stewart[1,*]

Research

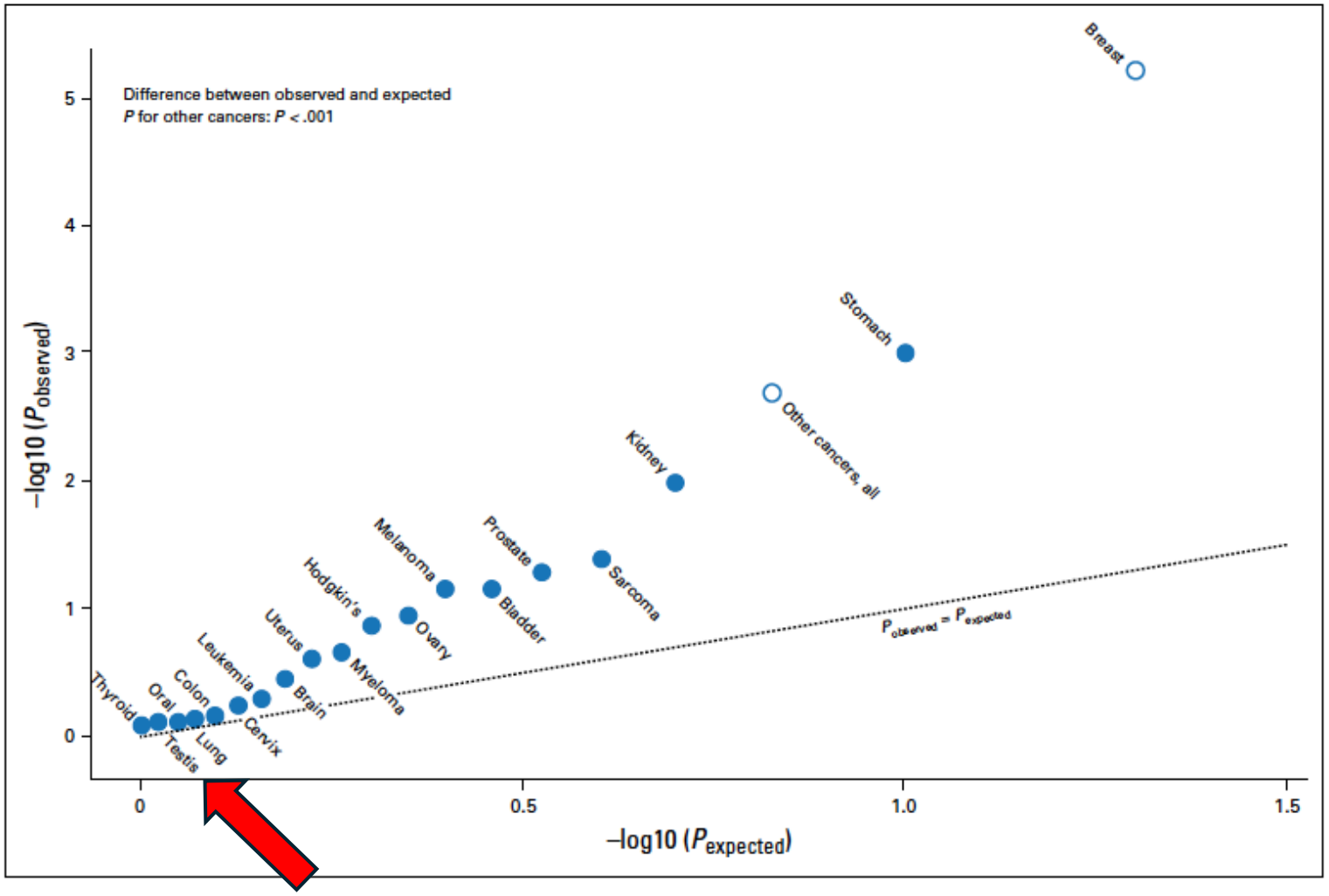
JAMA | Original Investigation

Estimated Prevalence and Clinical Manifestations of *UBA1* Variants Associated With VEXAS Syndrome in a Clinical Population

David B. Beck, MD, PhD; Dale L. Bodian, PhD; Vandan Shah, MD; Uyenlinh L. Mirshahi, PhD; Jung Kim, PhD; Yi Ding, MD, PhD; Samuel J. Magaziner, MPhil; Natasha T. Strande, PhD; Anna Cantor, MS; Jeremy S. Haley, MS; Adam Cook, MS; Wesley Hill; Alan L. Schwartz, MD, PhD; Peter C. Grayson, MD; Marcela A. Ferrada, MD; Daniel L. Kastner, MD, PhD; David J. Carey, PhD; Douglas R. Stewart, MD

CHEK2 is a low-to-moderate risk multi-tumor-predisposition gene

Risk of individual cancer types in *CHEK2* 1100delC heterozygotes



86,975 people from the Copenhagen General Population Study

Recruited 2003 – 2010

Linked to Danish Cancer Registry (1943 – 2011)

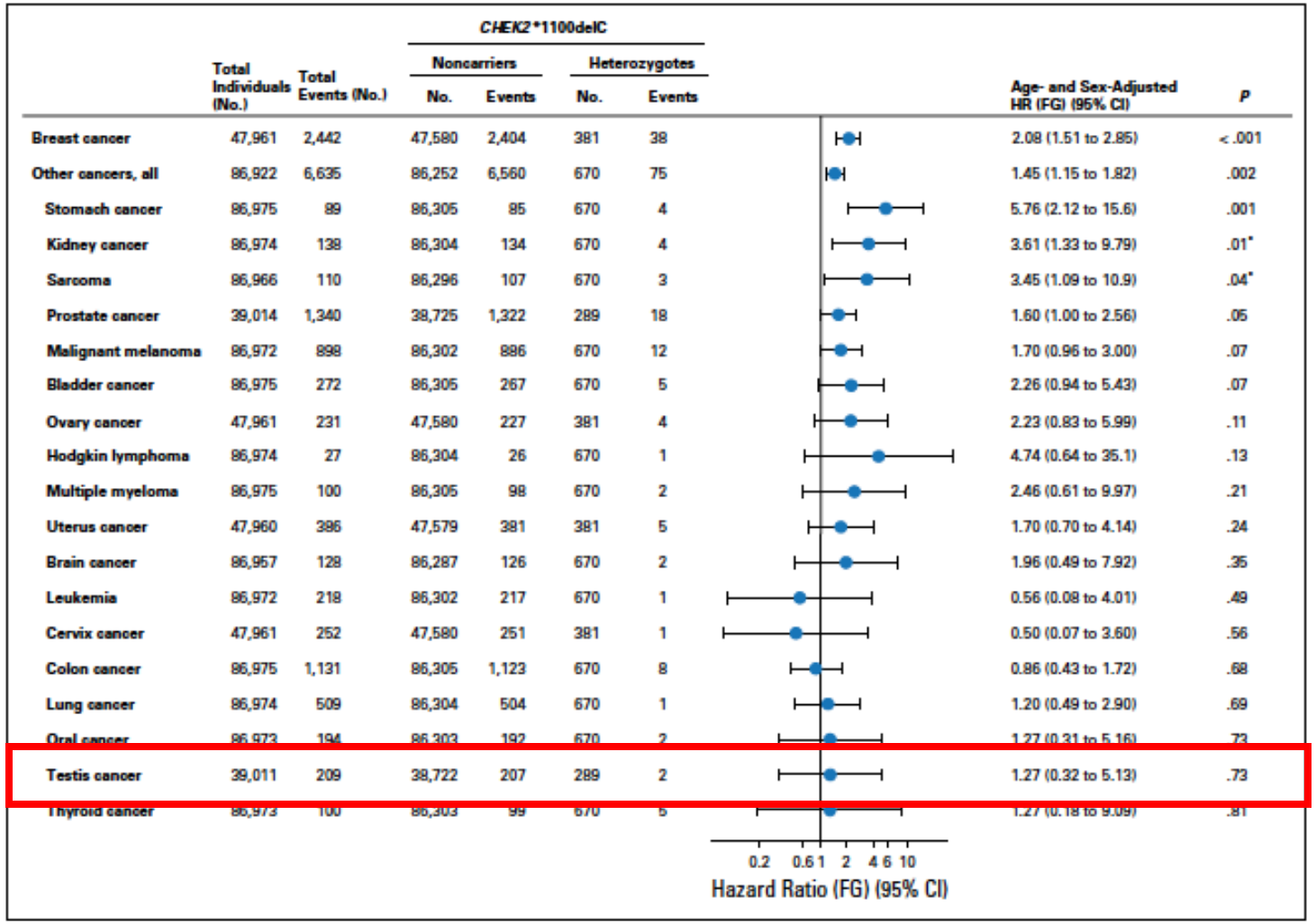
Open circles: *a priori* hypothesized cancers

Closed circles: exploratory analyses

Näslund-Koch *JCO* 2016

CHEK2 is a low-to-moderate risk multi-tumor-predisposition gene

Risk of individual cancer types in *CHEK2* 1100delC heterozygotes



	Total Individuals (No.)	Total Events (No.)	*CHEK2**1100delC Noncarriers No.	Noncarriers Events	Heterozygotes No.	Heterozygotes Events	Age- and Sex-Adjusted HR (FG) (95% CI)	*P*
Breast cancer	47,961	2,442	47,580	2,404	381	38	2.08 (1.51 to 2.85)	< .001
Other cancers, all	86,922	6,635	86,252	6,560	670	75	1.45 (1.15 to 1.82)	.002
Stomach cancer	86,975	89	86,305	85	670	4	5.76 (2.12 to 15.6)	.001
Kidney cancer	86,974	138	86,304	134	670	4	3.61 (1.33 to 9.79)	.01*
Sarcoma	86,966	110	86,296	107	670	3	3.45 (1.09 to 10.9)	.04*
Prostate cancer	39,014	1,340	38,725	1,322	289	18	1.60 (1.00 to 2.56)	.05
Malignant melanoma	86,972	898	86,302	886	670	12	1.70 (0.96 to 3.00)	.07
Bladder cancer	86,975	272	86,305	267	670	5	2.26 (0.94 to 5.43)	.07
Ovary cancer	47,961	231	47,580	227	381	4	2.23 (0.83 to 5.99)	.11
Hodgkin lymphoma	86,974	27	86,304	26	670	1	4.74 (0.64 to 35.1)	.13
Multiple myeloma	86,975	100	86,305	98	670	2	2.46 (0.61 to 9.97)	.21
Uterus cancer	47,960	386	47,579	381	381	5	1.70 (0.70 to 4.14)	.24
Brain cancer	86,957	128	86,287	126	670	2	1.96 (0.49 to 7.92)	.35
Leukemia	86,972	218	86,302	217	670	1	0.56 (0.08 to 4.01)	.49
Cervix cancer	47,961	252	47,580	251	381	1	0.50 (0.07 to 3.60)	.56
Colon cancer	86,975	1,131	86,305	1,123	670	8	0.86 (0.43 to 1.72)	.68
Lung cancer	86,974	509	86,304	504	670	1	1.20 (0.49 to 2.90)	.69
Oral cancer	86,973	194	86,303	192	670	2	1.27 (0.31 to 5.16)	.73
Testis cancer	39,011	209	38,722	207	289	2	1.27 (0.32 to 5.13)	.73
Thyroid cancer	86,973	100	86,303	99	670	1	1.27 (0.18 to 9.09)	.81

Näslund-Koch *JCO* 2016

Genomic ascertainment of *CHEK2*-related cancer predisposition

Sun Young Kim, Jung Kim, Mark Ramos, Jeremy Haley, Diane Smelser, H. Shanker Rao, Uyenlinh L. Mirshahi, Geisinger-Regeneron DiscovEHR Collaboration, Barry I. Graubard, Hormuzd A. Katki, David Carey, Douglas R. Stewart

doi: https://doi.org/10.1101/2024.08.07.24311613

Posted August 08, 2024.

Jung Kim, Ph.D.
Clinical Genetics Branch

Sun Young Kim, MD, Ph.D.
Cincinnati Children's

https://medrxiv.org/cgi/content/short/2024.08.07.24311613v1

CHEK2 heterozygotes (cases) and controls

Genomic ascertainment of cases and controls

Geisinger

- "Goldilocks" build; ABHet: 0.2-0.8; GQ>30; Depth > 5
- n= 167,050; age>18 yrs; mean 56.6 yrs
- *CHEK2* P/LP heterozygotes: 3,153
- 5 bi-alleleics with 2 common variants excluded
- Controls: individuals without *CHEK2* variation or B/LB: 152,662

UK Biobank

- Data-field 23157
- n=469,681; age>18 yrs; mean 56.5 yrs
- *CHEK2* P/LP heterozygotes: 3,232
- Includes 8 bi-alleleics but none with 2 common variants; not excluded
- Controls: individuals without *CHEK2* variation or B/LB: 305,330

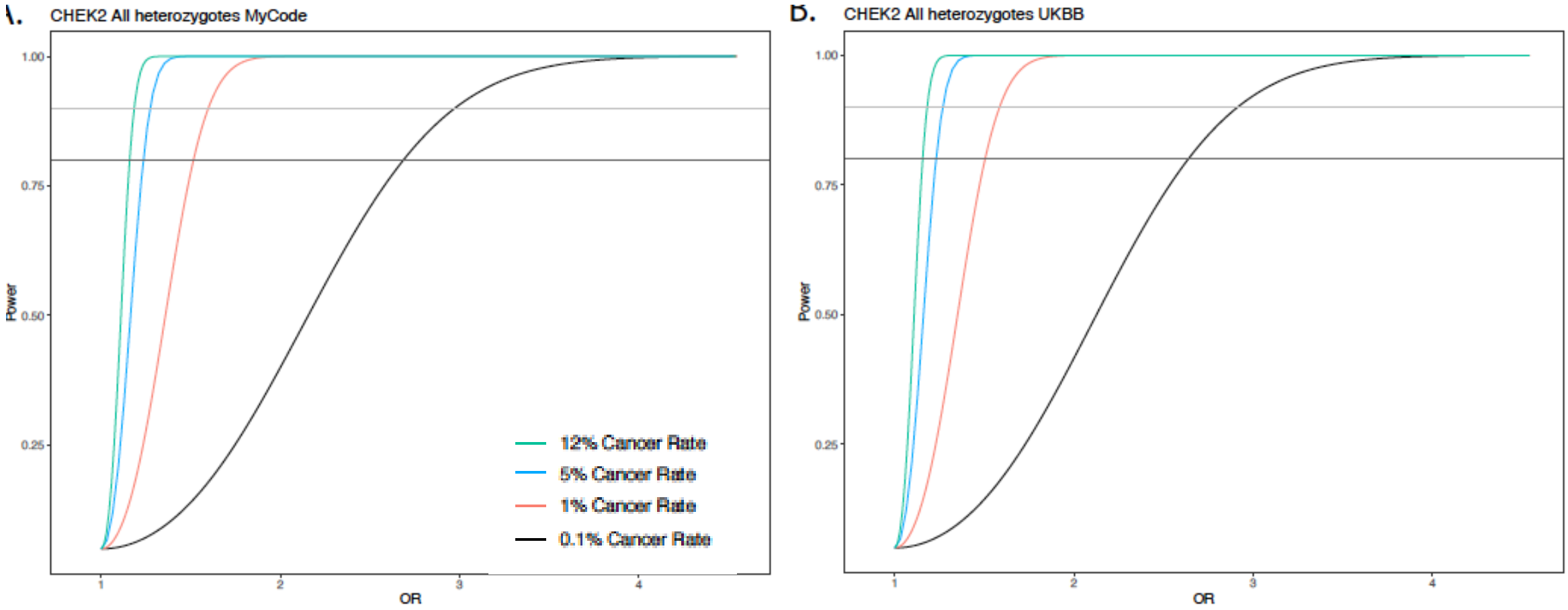
Kim, Kim *et al*. *Submitted*.

Prevalence of All, pathogenic truncating variants (PTV) and pathogenic missense variants (PMV) in *CHEK2* in adult heterozygotes in UK Biobank and Geisinger MyCode.

Cohort	Individuals/Prevalence (95%CI)	All *CHEK2* P/LP Variants	Pathogenic Truncating Variants (PTV)	Pathogenic Missense Variants (PMV)
UK Biobank – related and unrelated (n=469,765)	Number of individuals	3,232	1,847	1,290
	Prevalence	1/145 (1/140 – 1/150)	1/254 (1/243 – 1/266)	1/364 (1/344 – 1/384)
UK Biobank – unrelated (n=437,645)	Number of individuals	3,171	1,825	1,268
	Prevalence	1/138 (1/133-1/142)	1/239 (1/229-1/251)	1/345 (1/326-1/364)
MyCode – related and unrelated (n=167,050)	Number of individuals	3,153	913	2,221
	Prevalence	1/52 (1/51 – 1/54)	1/183 (1/171 – 1/195)	1/75 (1/72 – 1/78)
MyCode – unrelated (n=109,730)	Number of individuals	2,489	728	1,751
	Prevalence	1/43 (1/41 – 1/44)	1/150 (1/140-1/162)	1/62 (1/59-1/65)

Kim, Kim *et al*. *Submitted*.

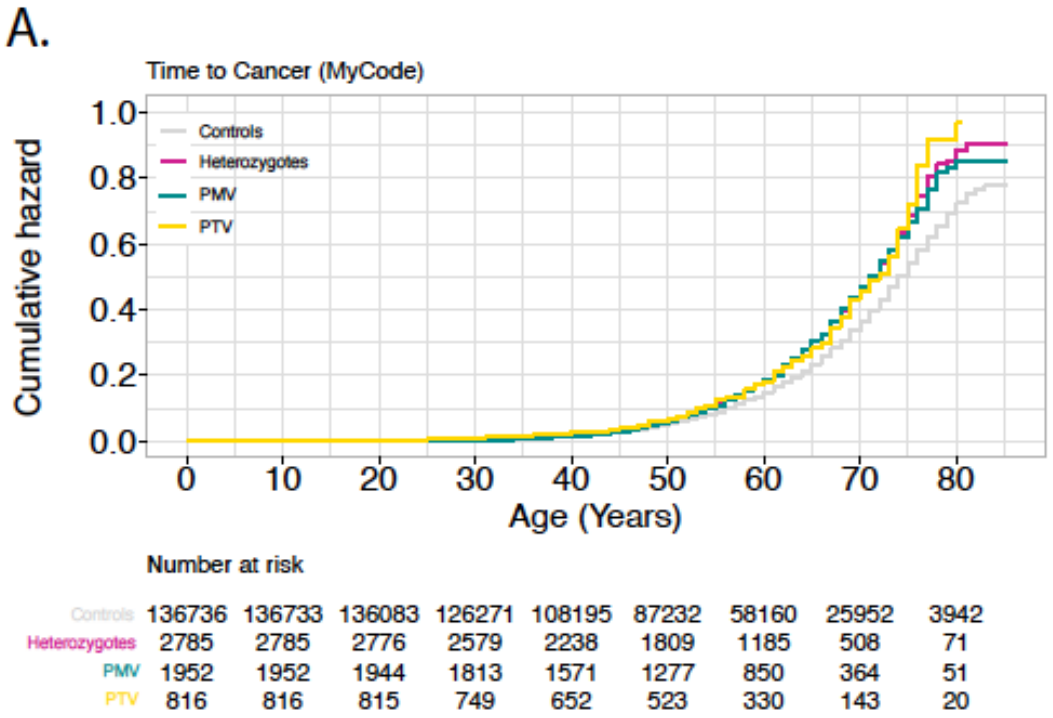
Power as a function of risk (odds ratio) in MyCode (left) and UK Biobank (right) for a range of cancer rates

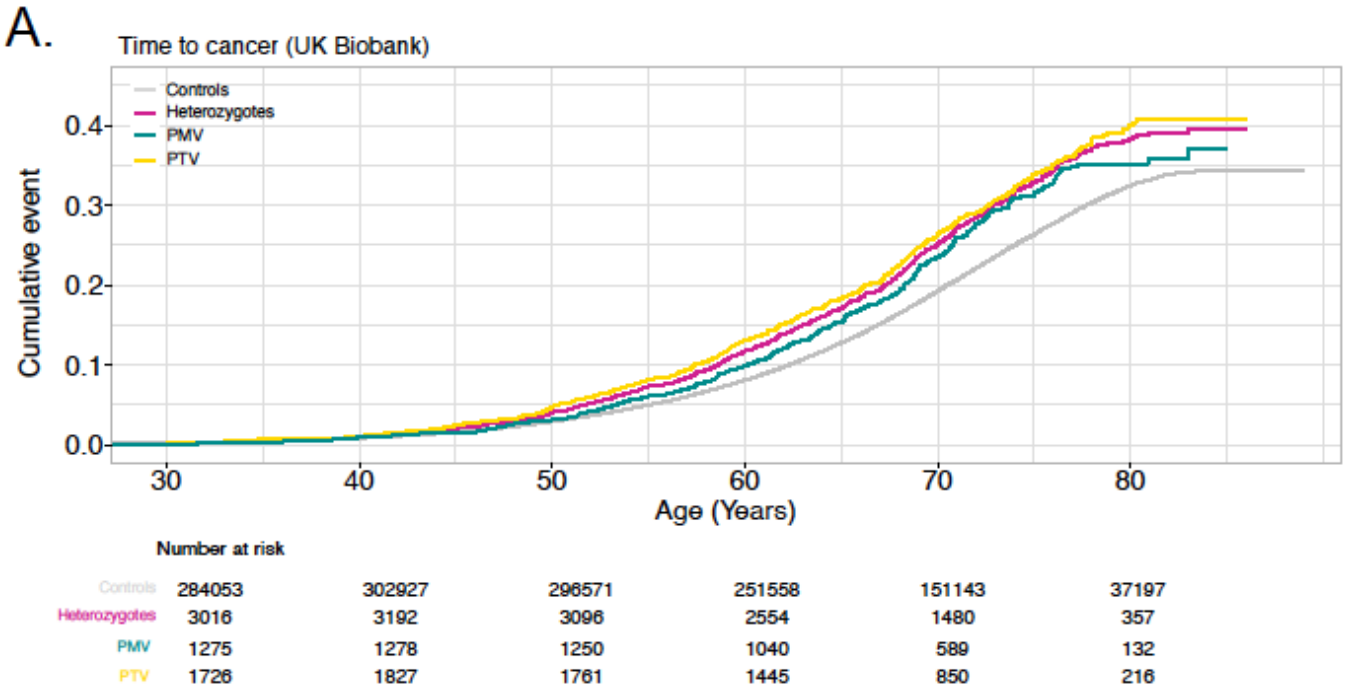


Dark gray line: 80% power; light gray line: 90% power

Kim, Kim *et al*. *Submitted*.

Age-dependent penetrance of pathogenic *CHEK2* variants for all cancers. Left: Time-to-cancer (penetrance) in Geisinger MyCode; **Right**: Time-to-cancer (penetrance) in UK Biobank



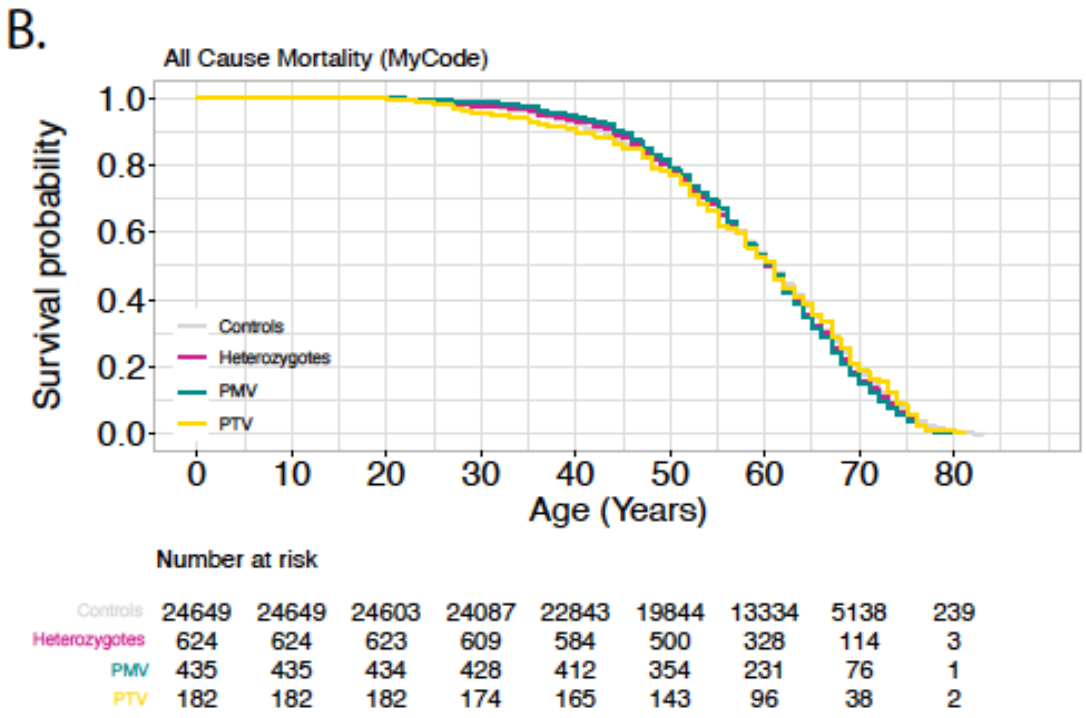


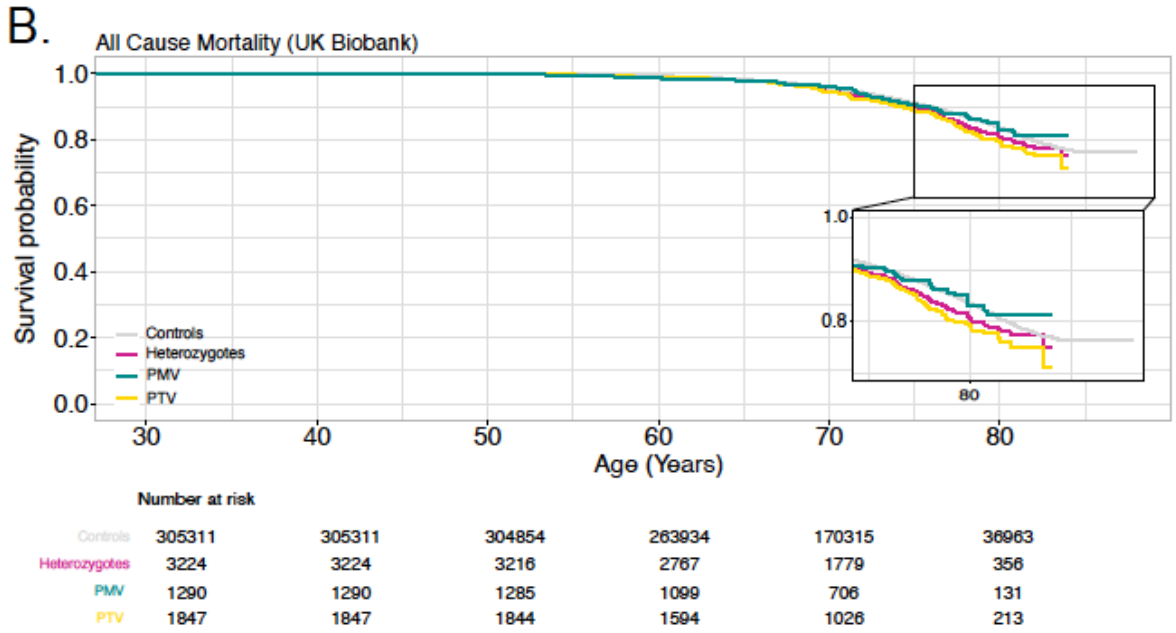
MyCode (adjusted HR: 1.26 [95%CI 1.16-1.37], *P*-value: 3.2×10^{-8})
(No significant difference between PMV and PTV)

UKBB (adjusted HR 1.31 [95%CI 1.24-1.40], *P*-value: 2×10^{-16})
(No significant difference between PMV and PTV)

Kim, Kim *et al*. *Submitted*.

Age-dependent penetrance of pathogenic *CHEK2* variants for all-cause mortality. Left: All-cause mortality in Geisinger MyCode; Right: All-cause mortality in UK Biobank





No significant differences in Geisinger MyCode
adjusted HR 1.09 [95%CI 0.96-1.24], *P*-value: 0.20)
(No significant difference between PMV and PTV)

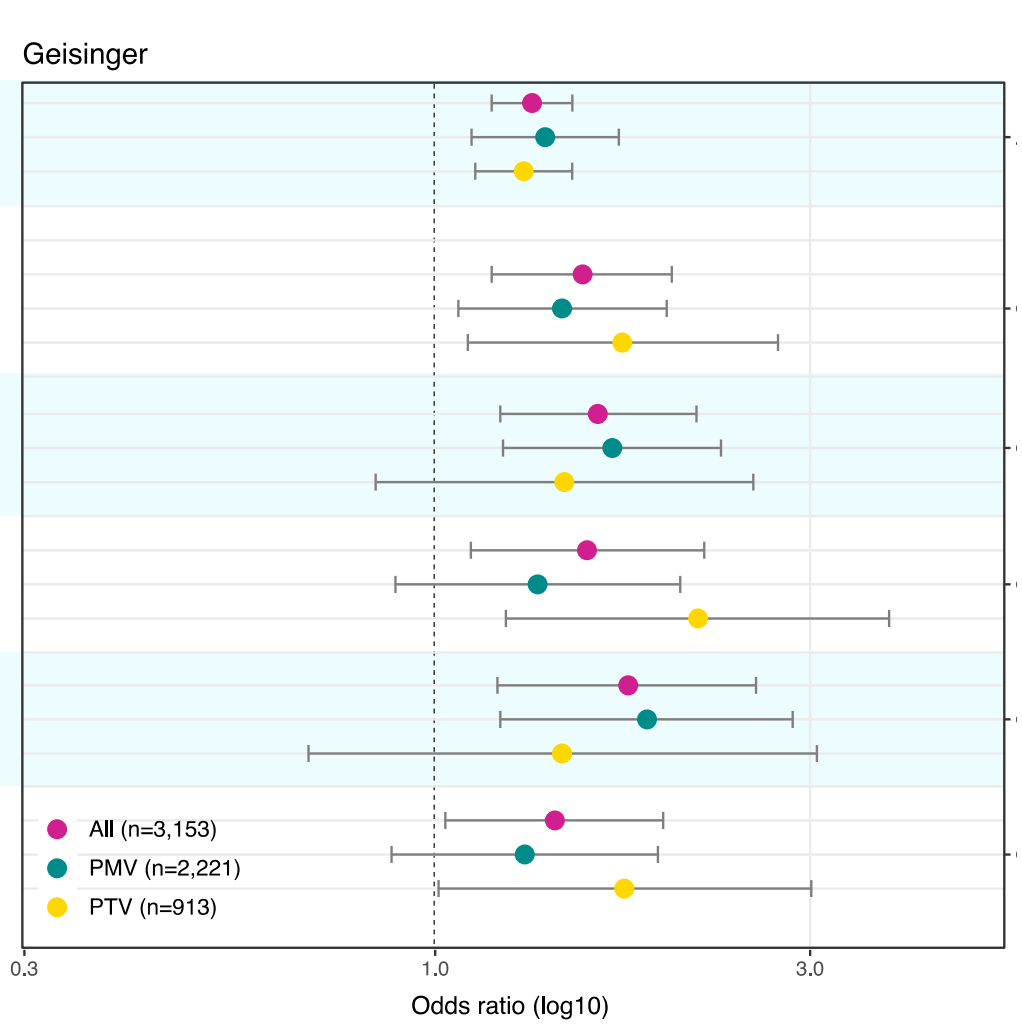
Significantly increased in All heterozygotes in UKBB
adjusted HR 1.21 [95%CI 1.08-1.37], P-value: 1.51×10^{-3}
(No significant difference between PMV and PTV)

Kim, Kim *et al*. *Submitted*.

Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for organ system groupings of cancer ICD codes with a significant excess of risk in Geisinger MyCode

A.

Controls(%)	Heterozygotes(%)	OR [95% CI]	Adjusted SAIGE p–value	
34,160(22.40)	864(27.40)	1.33[1.18–1.49]	2.15E–10	All cancers
34,160(22.40)	609(27.40)	1.38[1.11–1.71]	2.84E–07	
34,160(22.40)	247(27.1)	1.30[1.12–1.49]	1.09E–04	
5,004(3.28)	156(4.95)	1.54[1.18–2.00]	2.45E–05	C50: Breast
5,004(3.28)	105(4.73)	1.45[1.07–1.97]	1.05E–02	
5,004(3.28)	49(5.37)	1.73[1.10–2.73]	0.01	
4,443(7.49)	142(11.66)	1.61[1.21–2.15]	2.05E–04	C60-C63: Male genital
4,443(7.49)	106(12.08)	1.68[1.22–2.31]	6.90E–04	
4,443(7.49)	36(10.74)	1.46[0.84–2.54]	0.96	
2,632(1.72)	84(2.66)	1.56[1.11–2.2]	3.76E–03	C64-C68: Urinary tract
2,632(1.72)	53(2.39)	1.35[0.89–2.05]	0.7	
2,632(1.72)	31(3.40)	2.16[1.23–3.78]	0.03	
1,840(1.21)	66(2.09)	1.76[1.2–2.56]	9.11E–04	C73-C75: Thyroid/endocrine
1,840(1.21)	49(2.21)	1.86[1.21–2.85]	1.11E–03	
1,840(1.21)	16(1.75)	1.45[0.69–3.06]	1	
3,107(2.04)	91(2.88)	1.42[1.03–1.95]	0.04	C81-C96: Lymphoid/heme
3,107(2.04)	60(2.70)	1.3[0.88–1.92]	1	
3,107(2.04)	31(3.40)	1.74[1.01–3.01]	0.06	

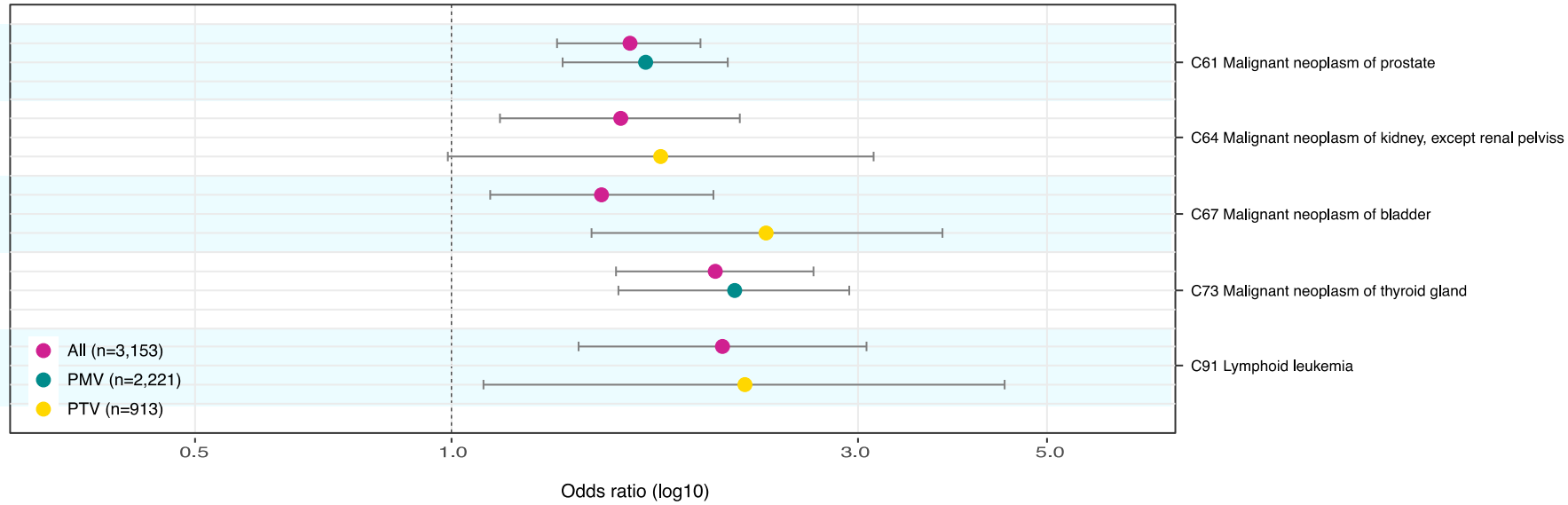


Kim, Kim *et al*. *Submitted*.

Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for specific cancers in the organ system groupings of cancer ICD codes with a significant excess of risk in Geisinger MyCode

A.

Controls(%)	Heterozygotes(%)	OR [95% CI]	Adjusted SAIGE p–value	MyCode
4222(7.12)	135(11.09)	1.62[1.27–2.07]	1.04E–04	C61 Malignant neoplasm of prostate
4222(7.12)	101(11.51)	1.69[1.27–2.24]	4.03E–04	
4222(7.12)	34(10.14)	NA	NA	
1202(0.79)	39(1.24)	1.58[1.03–2.41]	0.02	C64 Malignant neoplasm of kidney, except renal pelviss
1202(0.79)	27(1.22)	NA	NA	
1202(0.79)	12(1.31)	1.76[0.83–3.75]	0.28	
1477(0.96)	46(1.46)	1.50[1.01–2.23]	0.1	C67 Malignant neoplasm of bladder
1477(0.96)	27(1.22)	NA.	NA	
1477(0.96)	19(2.08)	2.34[1.254.38]	3.51E–03	
1370(0.90)	57(1.81)	2.04[1.48–2.82]	1.82E–05	C73 Malignant neoplasm of thyroid gland
1370(0.90)	42(1.89)	2.15[1.47–3.14]	6.60E–05	
1370(0.90)	14(1.53)	NA	NA	
627(0.41)	27(0.86)	2.08[1.17–3.69]	7.61E–03	C91 Lymphoid leukemia
627(0.41)	19(0.86)	NA	NA	
627(0.41)	8(0.88)	2.21[0.78–6.22]	0.53	

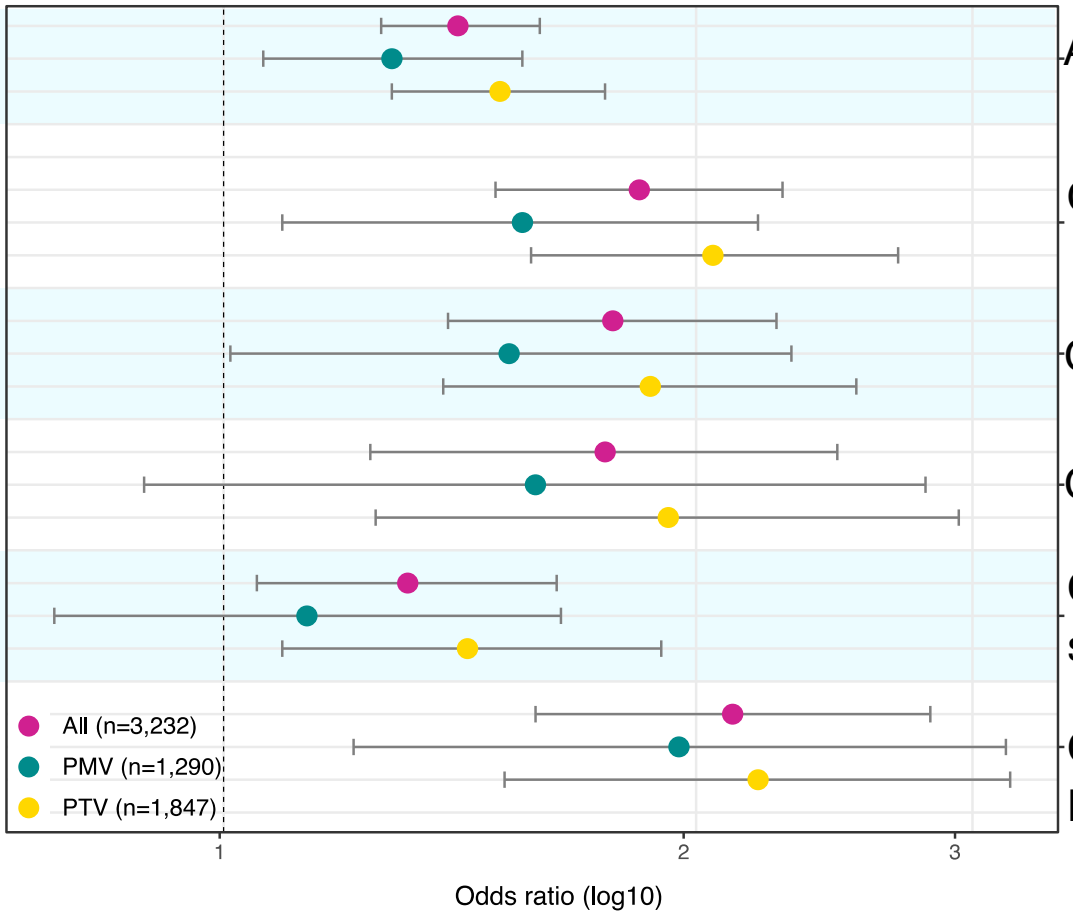


Kim, Kim *et al*. *Submitted*

Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for organ system groupings of cancer ICD codes with a significant excess of risk in UK Biobank

B.

Controls(%)	Heterozygotes(%)	OR [95% CI]	Adjusted SAIGE p–value	
69,836(22.87)	934(28.90)	1.41[1.26–1.59]	6.05E–15	All cancers
69,836(22.87)	344(26.67)	1.28[1.06–1.55]	2.19E–04	
69,836(22.87)	562(30.23)	1.50[1.28–1.75]	9.56E–12	
12,439(4.07)	227(7.02)	1.84[1.49–2.27]	2.47E–12	C50: Breast
12,439(4.07)	80(6.20)	1.55[1.09–2.19]	1.41E–02	
12,439(4.07)	139(7.53)	2.05[1.57–2.69]	4.49E–10	
10,557(3.46)	179(5.53)	1.77[1.39–2.25]	3.64E–09	C60-C63: Male genital
10,557(3.46)	58(4.50)	1.52[1.01–2.30]	0.14	
10,557(3.46)	113(6.12)	1.87[1.38–2.53]	4.13E–07	
4,351(1.43)	77(2.38)	1.75[1.24–2.46]	1.01E–04	C64-C68: Urinary tract
4,351(1.43)	27(2.09)	1.58[0.89–2.80]	0.84	
4,351(1.43)	49(2.65)	1.92[1.25–2.94]	2.68E–04	
14,339(4.70)	192(5.94)	1.31[1.05–1.63]	4.68E–03	C76-C79: Ill-defined, secondary, unspecifie
14,339(4.70)	66(5.12)	1.13[0.78–1.64]	1	
14,339(4.70)	120(6.50)	1.43[1.09–1.90]	2.95E–03	
5,030(1.65)	108(3.34)	2.11[1.58–2.82]	3.45E–09	C81-C96: Lymphoid/heme
5,030(1.65)	39(3.02)	1.95[1.21–3.15]	4.70E–03	
5,030(1.65)	65(3.52)	2.19[1.51–3.17]	8.74E–07	

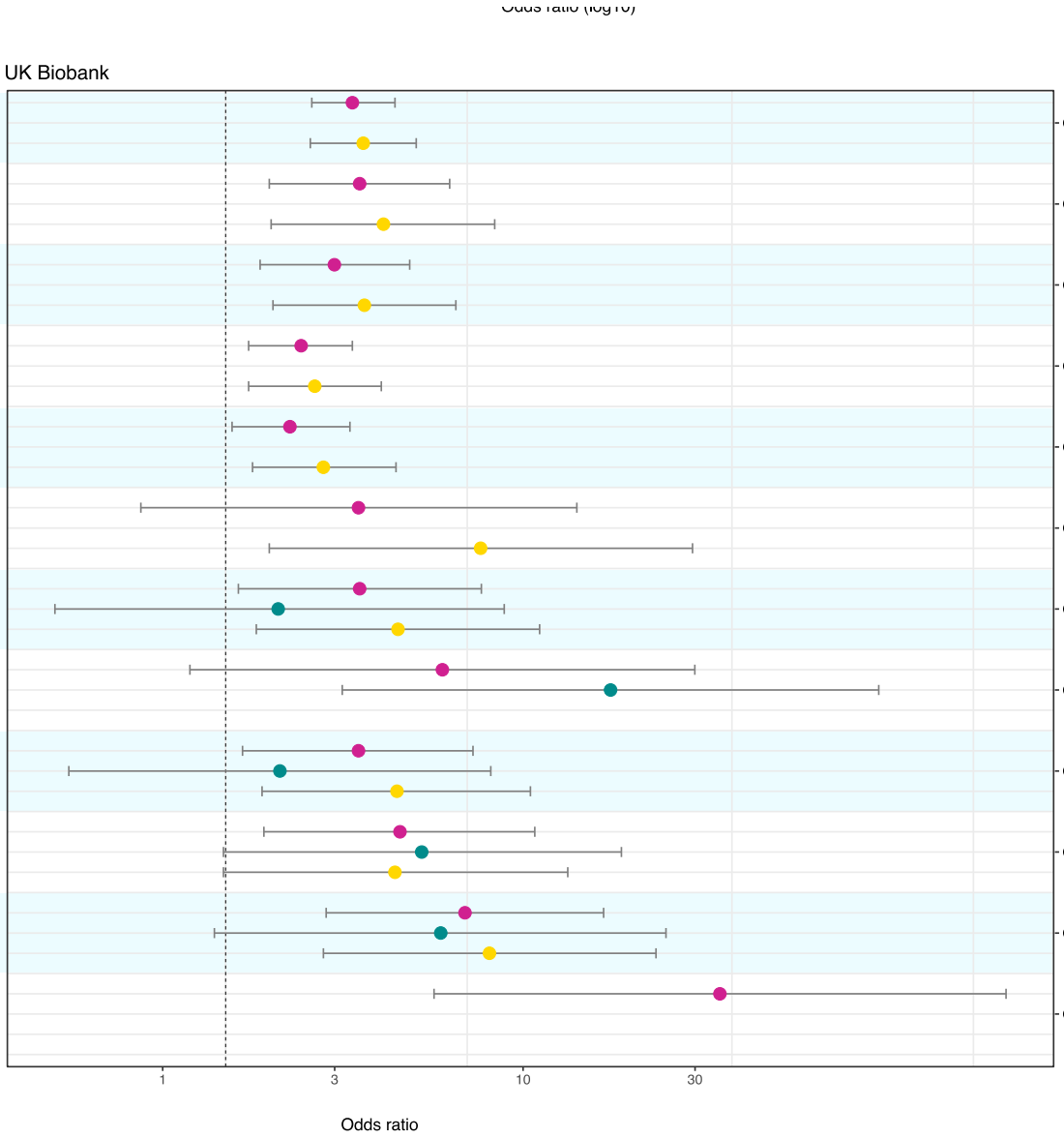


Kim, Kim *et al*. *Submitted*.

Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for specific cancers in the organ system groupings of cancer ICD codes with a significant excess of risk in UK Biobank

B.

Controls(%)	Heterozygotes(%)	OR [95% CI]	Adjusted SAIGE p–value	UK Biobank
9,978(3.27)	170(5.27)	1.78[1.48–2.16]	6.68E–10	C61 Malignant neoplasm of prostate
9,978(3.27)	55(4.26)	NA	NA	
9,978(3.27)	107(5.79)	1.87[1.47–2.38]	1.27E–07	
1,635(0.54)	31(0.96)	1.84[1.22–2.77]	2.49E–03	C64 Malignant neoplasm of kidney, except renal pelvis
1,635(0.54)	11(0.85)	NA	NA.	
1,635(0.54)	20(1.08)	2.05[1.23–3.40]	4.28E–03	
2,707(0.89)	45(1.40)	1.64[1.17–2.31]	2.66E–03	C67 Malignant neoplasm of bladder
2,707(0.89)	15(1.16)	NA	NA	
2,707(0.89)	30(1.62)	1.88[1.24–2.85]	2.32E–03	
7,493(2.45)	109(3.38)	1.41[1.11–1.78]	9.77E–04	C77 Secondary and unspecified malignant neoplasm of lymph nodes
7,493(2.45)	40(3.10)	NA	NA	
7,493(2.45)	66(3.57)	1.50[1.11–2.03]	4.23E–03	
6,072(2.00)	84(2.61)	1.34[1.03–1.76]	0.01	C79 Secondary malignant neoplasm of other sites
6,072(2.00)	27(2.09)	NA	NA	
6,072(1.99)	56(3.03)	1.56[1.13–2.17]	4.40E–03	
414(0.14)	8(0.25)	1.83[0.68–4.94]	0.02	C81 Hodgkin's disease
414(0.14)	0(0)	NA	NA	
414(0.14)	8(0.43)	3.19[1.22–8.36]	8.40E–03	
1,363(0.45)	26(0.81)	1.84[1.06–3.20]	0.03	C83 Diffuse non–Hodgkin's lymphoma
1,363(0.45)	7(0.54)	1.27[0.46–3.55]	1	
1,362(0.45)	18(0.97)	2.19[1.15–4.17]	0.01	
216(0.07)	6(0.19)	2.68[0.85–8.45]	0.24	C84 Peripheral and cutaneous T–cell lymphomas
216(0.07)	5(0.39)	5.76[1.70–19.50]	0.02	
216(0.07)	1(0.05)	NA	NA	
1,533(0.50)	29(0.90)	1.83[1.08–3.08]	0.02	C85 Other and unspecified types of non–Hodgkin's lymphoma
1,533(0.50)	8(0.62)	1.28[0.49–3.34]	0.94	
1,533(0.50)	20(1.08)	2.18[1.18–4.00]	0.02	
921(0.30)	21(0.65)	2.21[1.19–4.08]	0.01	C91 Lymphoid leukaemia
921(0.30)	9(0.70)	2.44[0.99–6.05]	0.14	
921(0.30)	12(0.65)	2.16[0.99–4.74]	0.08	
650(0.21)	20(0.62)	2.97[1.58–5.58]	2.17E–04	C92 Myeloid leukaemia
650(0.21)	7(0.54)	2.66[0.95–7.41]	0.13	
650(0.21)	13(0.70)	3.32[1.56–7.08]	1.22E–03	
50(0.02)	5(0.16)	9.47[2.58–34.80]	1.91E–03	C96 Other and unspecified malignant neoplasms of lymphoid, haematopoietic and related tissue
50(0.02)	3(0.23)	NA	NA	
50(0.02)	2(0.11)	NA	NA	



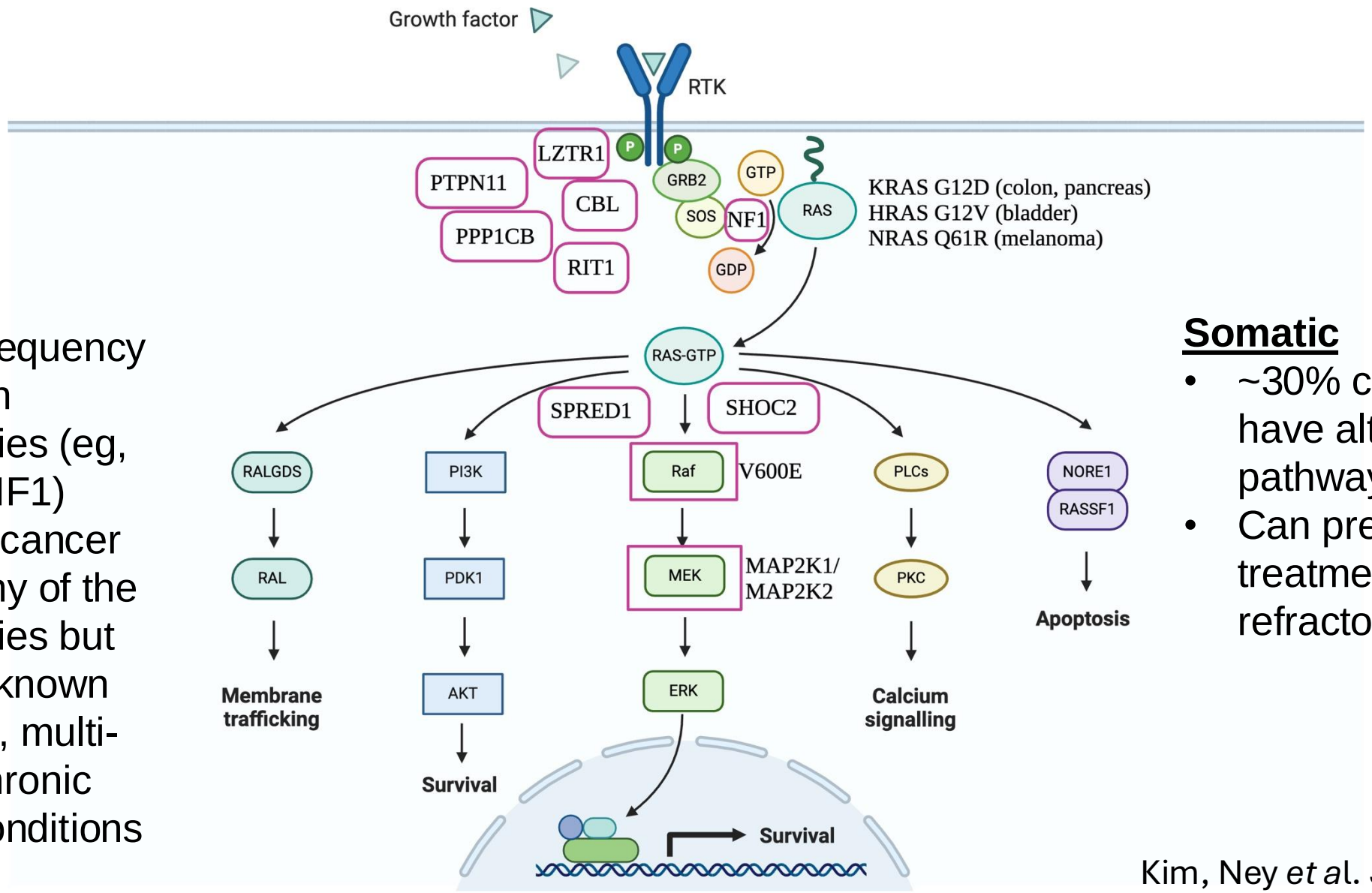
Kim, Kim *et al*. *Submitted*

Genomic ascertainment of *CHEK2* heterozygotes

- Relatedness-adjusted, Bonferroni-corrected genomic ascertainment of two population-based, exome-sequenced, EHR-linked cohorts
- High power to detect elevated risk (OR>2) in all but the rarest cancers
- Confirms the significantly increased risk for breast and prostate cancers (as well as all cancers, collectively)
- Observed risk tends to be even lower (OR<2) than previous estimates, especially for PTV
- In neither cohort was a significant excess risk for "malignant neoplasms of digestive organs" observed, despite numerous studies in which a modest excess risk has been reported
- Substantial evidence from both cohorts of significant increased risk for kidney cancer, bladder cancer and CLL (lymphoid leukemia).
- Significant excess of malignancies of thyroid and other endocrine tumors (C73-C75) was observed in MyCode but not UK Biobank
- For some rarer cancers (male breast, testicular) the two cohorts were likely underpowered for others (sarcoma, stomach) there may be both a power issue and a survival bias in ascertainment given the aggressive nature of these cancers

Kim, Kim *et a*l. *Submitted*

RAS/MAPK Pathway



Germline

- ~1:2000 frequency of common RASopathies (eg, Noonan, NF1)
- Increased cancer risk in many of the RASopathies but degree unknown
- Significant, multi-system, chronic medical conditions

Somatic

- ~30% cancers have altered RAS pathway
- Can predict treatment refractoriness

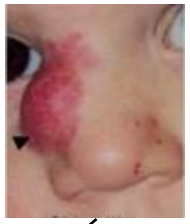
Kim, Ney *et al*. *Submitted*

The RASopathies

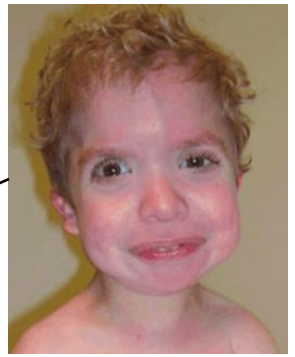
SYNGAP1

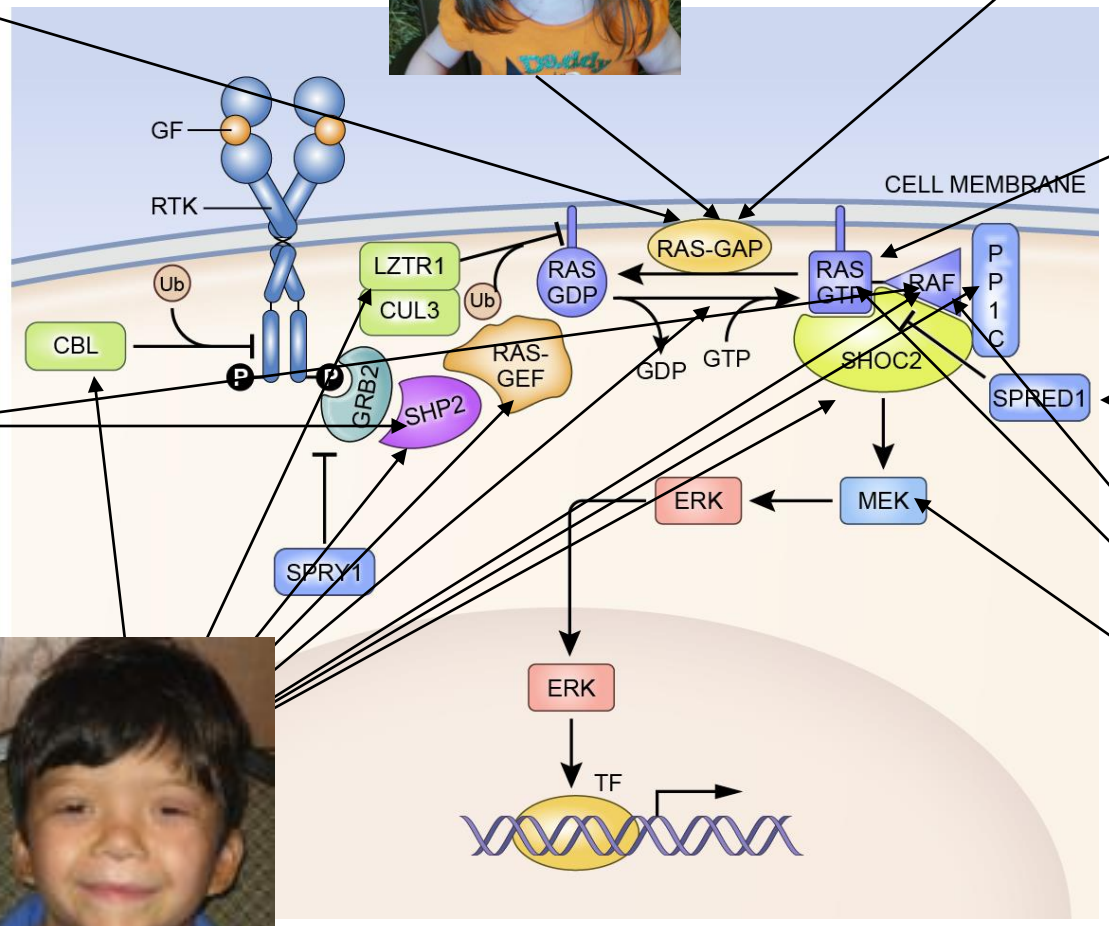
NF1

Cap-AV Malformation



Costello Syndrome





Noonan Syndrome
Multiple Lentigines

Legius Syndrome

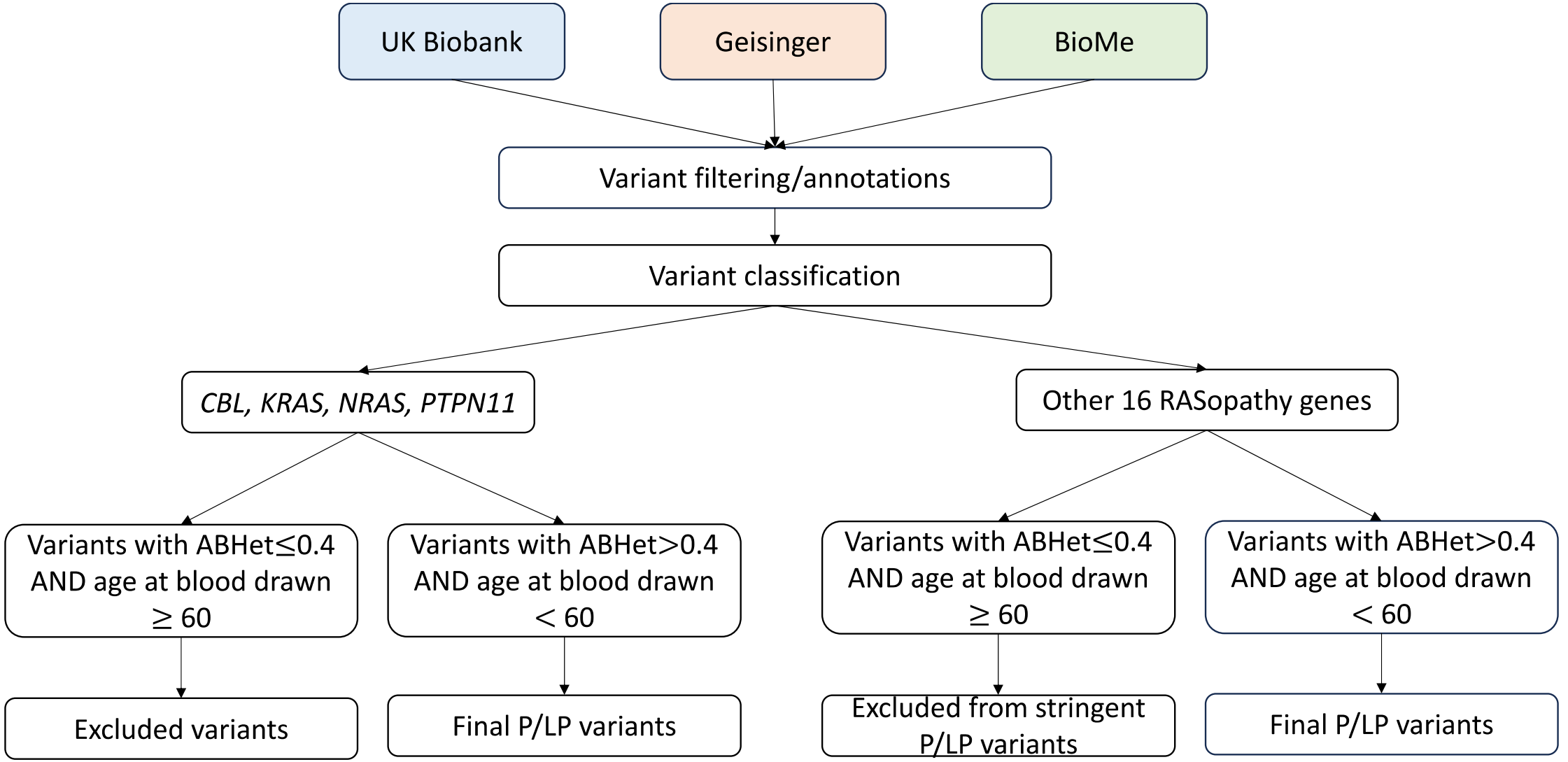
Noonan Syndrome

Cardiofaciocutaneous Syndrome

Genomic ascertainment of RASopathies

- Cancer risk well documented in childhood and adolescence
  - From phenotypic and family ascertatinment
  - Costello: bladder, rhabdomyosarcoma, neuroblastoma
  - High-risk variants in Noonan: JMML, rhabdomyosarcoma, neuroblastoma
- Cancer risk in CFC and Legius syndrome unclear throughout lifespan
- Despite case reports of cancer in adults with germline P/LP in RASopathy genes, cancer risk is unknown
- Interrogate the exome sequence of individuals in three large biobanks to quantify germline P/LP variant prevalence, cancer incidence, and survival of adults with non-NF1 RAS/MAPK genes

Flow diagram of variant classification and application of filtering to limit clonal hematopoiesis (CH) variants and develop a range of germline prevalence estimates.



Kim, Ney *et al*. *Submitted*

Frequency of RAS/MAPK variants in UKBiobank, Geisinger, and Bio*Me*

		UK Biobank		Geisinger MyCode		Mount Sinai BioMe	
		(469,618 exomes)		(167,050 exomes)		(30,129 exomes)	
Gene Syndrome		count	frequency	count	frequency	count	freq
CBL		14	1:33,544 (1:19,982–1:56,309)	8	1:20,881 (1:10,581–1:41,207)	0	-
CFC	stringent	9	1:52,179 (1:27,453–1:99,178)	3	1:55,683 (1:17,474–1:215,605)	0	-
		10	1:46,961 (1:25,509–1:86,453)	4	1:41,762 (1:15,185–1:130,367)	0	-
Noonan	stringent	141	1:3,330 (1:2,824–1:3,927)	68	1:2,456 (1:1,938–1:3,113)	15	1:2,008 (1:1,217–1:3,314)
		149	1:3,151 (1:2,684–1:3,700)	73	1:2,288 (1:1,820–1:2,876)	17	1:1,772 (1:1,106–1:2,838)
NSML		21	1:22,362 (1:14,627–1:34,188)	7	1:23,864 (1:11,560–1:49,264)	1	1:30,129 (1:4,639–1:577,181)
Noonan without NSML	stringent	120	1:3,913 (1:3,273–1:4,679)	61	1:2,738 (1:2,132–1:3,517)	14	1:2,152 (1:1,282–1:3,612)
		128	1:3,668 (1:3,086–1:4,361)	66	1:2,531 (1:1,989–1:3,219)	16	1:1,883 (1:1,159–1:3,058)
Legius (*SPRED1*)	stringent	24	1:19,567 (1:13,150–1:29,116)	3	1:55,683 (1:17,474–1:215,605)	0	-
		24	1:19,567 (1:13,150–1:29,116)	4	1:41,762 (1:15,185–1:130,367)	0	-
Costello		0	-	0	-	0	-

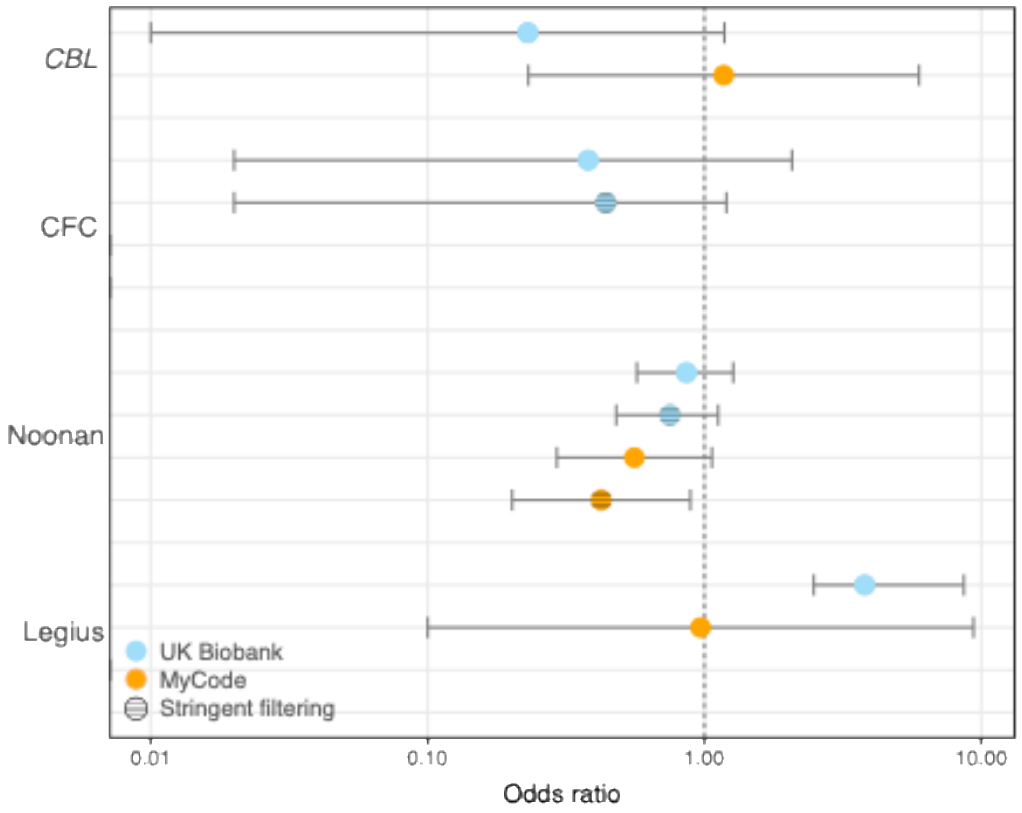
Noonan – *KRAS, MRAS, PTPN11, SOS1, SOS2, LZTR1, PPP1CB, RRAS1, RRAS2, SHOC2, RIT1, RAF1, NRAS, RASA2*

Cardiofaciocutaneous (CFC): *BRAF, MAP2K1, MAP2K2*

NSML: Noonan syndrome with multiple lentigines: select variants in *PTPN11*

Stringent: excludes all variants in genes with ABHet ≤ 0.4 and age at blood draw ≥ 60 years

Cancer prevalence calculated as Odds Ratio in individuals with germline Pathogenic/Likely Pathogenic variants in RASopathies versus controls in UKBB and Geisinger cohorts



	Controls(%)	Heterozygotes with cancer(%)	OR [95% CI]	p–value
	113,110(24.2)	1(7.14)	0.23[0.01–1.18]	0.16
	39,352(23.9)	2(25.0)	1.17[0.23–5.95]	0.16
	113,110(24.2)	1(10.0)	0.38[0.02–2.07]	0.37
	113,110(24.2)	1(11.1)	0.44[0.02–1.20]	0.45
	39,352(23.9)	0(0)	NA	NA
	39,352(23.9)	0(0)	NA	NA
	113,110(24.2)	32(21.5)	0.86[0.57–1.27]	0.47
	113,110(24.2)	27(19.1)	0.75[0.48–0.89]	0.18
	39,352(23.9)	11(15.1)	0.56[0.29–1.06]	0.08
	39,352(23.9)	8(11.8)	0.42[0.20–0.89]	0.02
	113,110(24.2)	13(54.2)	3.79[2.48–8.64]	1.20E–03
	39,352(23.9)	1(25.0)	0.97[0.10–9.37]	0.98
	39,352(23.9)	0(0)	NA	NA

CBL syndrome: *CBL*

Cardiofaciocutaneous (CFC): *BRAF, MAP2K1, MAP2K2*

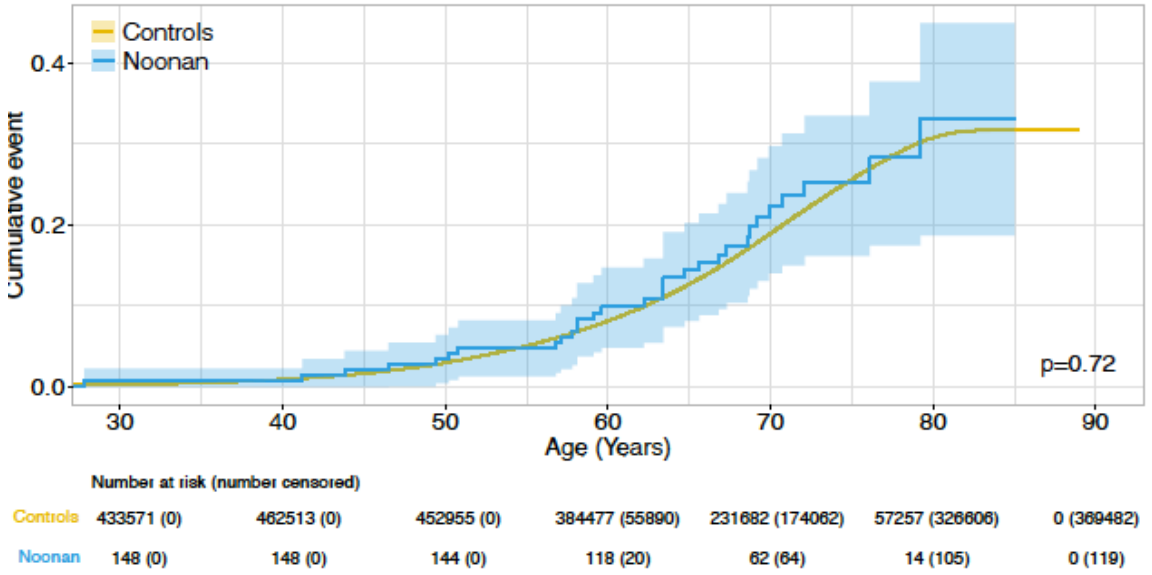
Noonan: *KRAS, MRAS, PTPN11, SOS1, SOS2, LZTR1, PPP1CB, RRAS1, RRAS2, SHOC2, RIT1, RAF1, NRAS, RASA2*

Legius: *SPRED1*

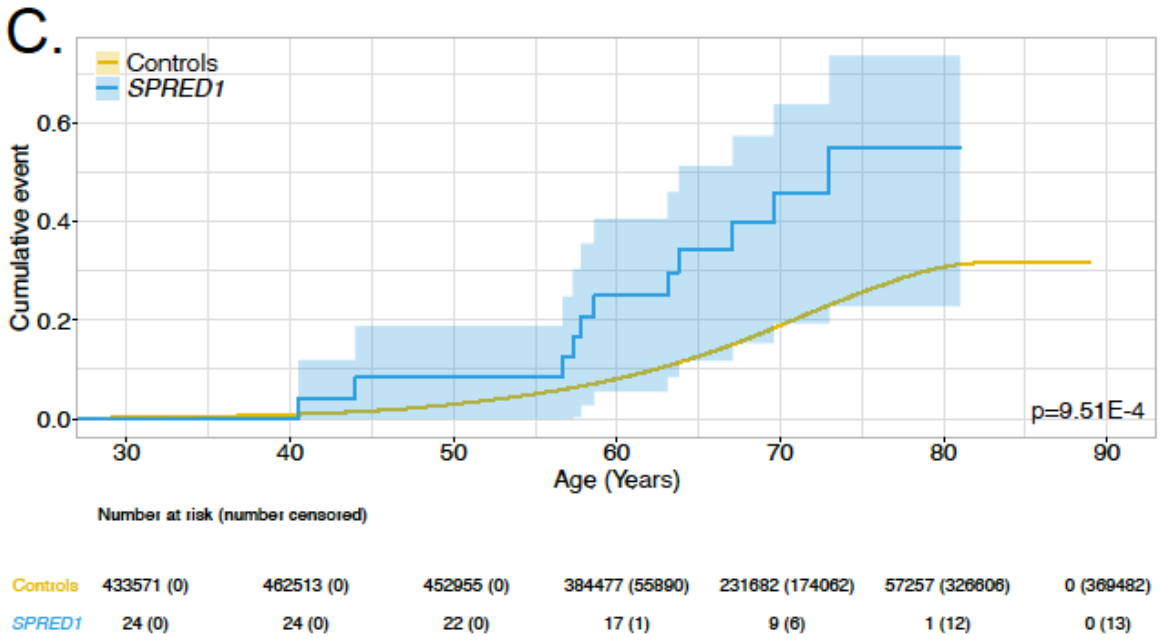
Stringent: excludes all variants in genes with ABHet ≤ 0.4 and age at blood draw ≥ 60 years

Time to cancer in UK Biobank in Noonan-associated genes (panel A) and in individuals with Pathogenic/Likely Pathogenic variants in *SPRED1* (panel C)

A.

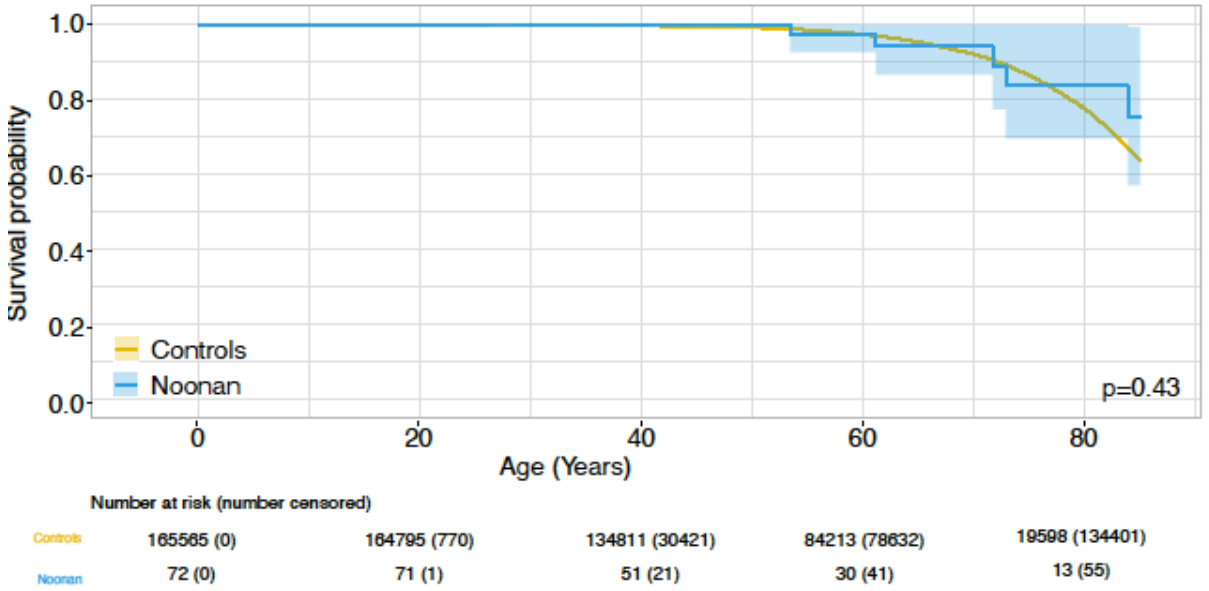


C.

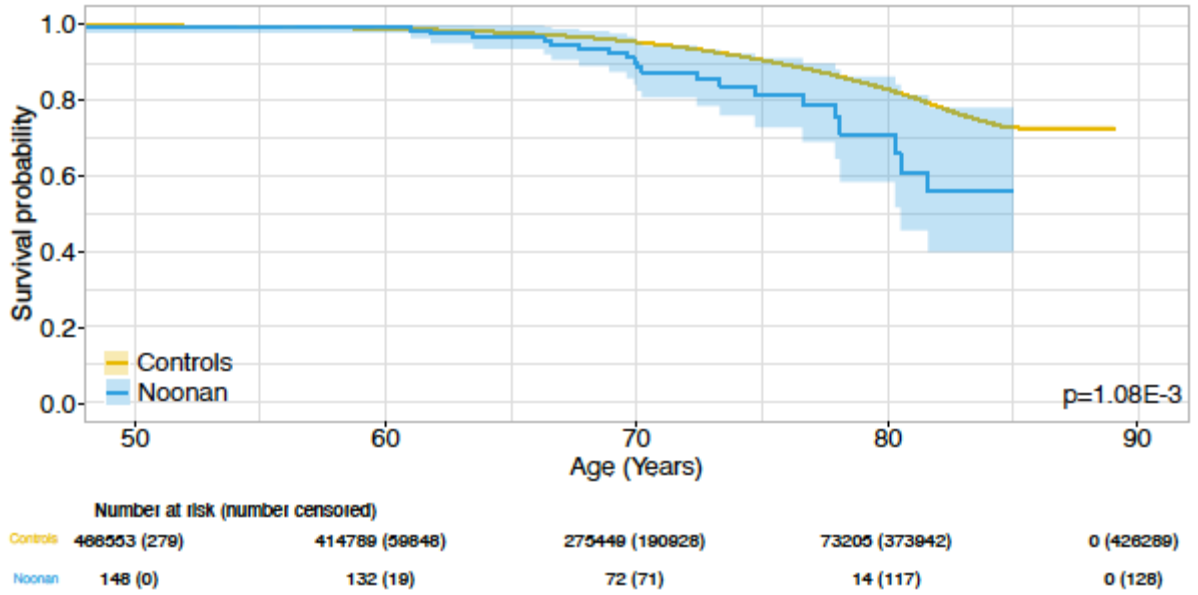


Time-dependent survival is not significantly different in Noonan-heterozygotes vs. controls in Geisinger MyCode (panel A) but is less favorable vs controls in UK Biobank (panel C)

A.



C.



Using Genomic Ascertainment to Explore Prevalence and Cancer Risk in Adult Individuals with Pathogenic and Likely Pathogenic Germline Variants in RASopathy Genes

- P/LP variants in Noonan syndrome-associated genes were the most common
- P/LP variants in Noonan syndrome-associated genes were not associated with an increased cancer risk in adults
- In UK Biobank, P/LP variants in *SPRED1* were associated with a 4-fold higher risk of cancer compared to controls in adults and had earlier cancer onset
- In UK Biobank, P/LP variants in Noonan syndrome-associated genes were correlated with increased all-cause mortality and cancer-related mortality

What are the consequences of genomic ascertainment?

- *Prevalence* of pathogenic/likely pathogenic (P/LP)* variants is greater than previously estimated

- *Penetrance* (*risk* from a P/LP variant) may not be as high as previously estimated

- *Phenotype* is different (may be less severe, broader)

*Clinically actionable germline variation classified by ACMG/AMP rules (Richards *et al Genetics in Medicine* 2015)

Lessons learned from the genome-first approach (so far)– part I

- *Genome-first approach holds enormous promise, however…*

- Complements “phenotype-first” approach
- Manual EHR review is messy, incomplete and labor-intensive
  - Often query of ICD codes tells you what you need to know
- Large cohort x rare disease = modest numbers
  - More sequencing: DCEG Connect, NIH All of Us, UK Biobank …
  - Usefulness of phenotype-first cohorts
- Characteristics of cohorts matter and bring their own biases
  - Health system vs. healthier volunteer
- Outcome studies are easier to do than etiology studies

Lessons learned from the genome-first approach (so far)– part II

- *Genome-first approach holds enormous promise, however…*

- Large number of matched controls a blessing and a curse (inflated p-values)
  - Work with a good biostatistician
- Variant interpretation is relatively easier
  - We focus on ACMG/AMP classification of pathogenicity (for now)
  - Work with a good variant scientist
- Phenotype work is relatively harder
  - Pick your phenotypes with care and keep simple: height, cancer registry, blood glucose
  - Work with a clinical bioinformatician who knows ICD coding and phecodes
  - Phecodes as a way to simplify use of ICD codes
- Medical coding is an art and science
  - Multiplicity/redundancy of codes for the same thing
  - Institutional coding cultures
  - Awareness of diagnoses: breast/colon cancer family hx vs renal cancer family hx

What's next?

- Analyze larger and larger cohorts
  - All of Us (NIH)
    - Goal of 1 million participants
    - Reflects diversity of the US circa 2024
- Analyze genome (not "just" exome) data
  - Wide variety of pathogenic variants
  - All of Us releases genome data – available on ~250,000 people now
  - UK Biobank release of genome data on 500,000 people
- Recruitment using genomic ascertainment
  - Reverse Phenotyping Core (NHGRI/Les Biesecker)
  - All of Us (NIH/Josh Denny)
- Incorporate findings into surveillance and variant interpretation guidelines
- Overall goal: improve "cancer interception": grape vs. grapefruit

Practicalities using MyCode data

- Use: genome-first, GWAS, ExWAS (rare-variant association)…
- Available data: ~233K exomes (arrays) with linked demographics, ICD codes, labs, imaging, chart review, pathology (samples and reports), medications, visit type, orthogonal sequencing
- Includes ~9000 pediatric exomes
- Access: though Jung and Doug
- Scheduled calls
- Proposal form
- Cost (CGB)
- Logistics of running analyses