Appendices for

Title: A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records

Authors: Tao Zheng[1,2], Wei Xie[3], Liling Xu[2], Xiaoying He[4], Ya Zhang[1], Mingrong You[5], Gong Yang[5], You Chen[6]

Author Affiliations:
[1]Institute of Image Communication and Networking, Shanghai Jiao Tong University, Shanghai, China
[2] Tongren Hospital Shanghai Jiao Tong University, Shanghai, China
[3] Department of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN, USA
[4] Department of Endocrinology, the First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China
[5] Division of Epidemiology, Vanderbilt University, Nashville, TN, USA
[6] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

To Whom Correspondence Should be Addressed:

You Chen, Ph.D.
2525 West End Ave, Suite 1475
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN 37203 USA
Email: you.chen@vanderbilt.edu

Keywords: Electronic Health Records; Type 2 Diabetes, Data Mining, Feature Construction,

Machine Learning

Appendix A: A list of 110 constructed features

The constructed 110 features across seven sources are listed in Table A1. For every source, we design specific features covering diagnosis codes (ICD-10 codes E11. ***), diagnosis notes (positive notes and negative notes as shown in Table A3 of Appendix C), self-report notes (persistent hunger, polyuria, and polydipsia), medications (traditional Chinese medicine and western medicine), plasma glucose test (venous and peripheral) and HbA1C test.

Table A1. The constructed 110 features coming from seven sources.

Source	category	Feature
Demographic information		f1: De-identification ID of a subject
		f2: An integer number representing age
		f3: Gender
Communication report	Self-reporting note	f4: Number of times a subject reporting body weight loss
		f5: Number of times a subject reporting persistent hunger
		f6: Number of times a subject reporting polyuria
		f7: Number of times a subject reporting polydipsia
		f8: Number of times a subject reporting prescribed diabetes medicine
		f9: Number of returning visits for diabetes
	Diagnosis code	f10: Number of times type 2 diabetes codes were assigned
		f11: Number of times diabetes codes were assigned, but the type of diabetes is not specified
		f12: Number of times diabetic retinopathy codes were assigned
		f13: Number of times diabetic neuropathy codes were assigned
		f14: Number of times diabetic eye disease codes were assigned
		f15: Number of times diabetic kidney disease codes were assigned
		f16: Number of times diabetic cerebral vascular disease codes were assigned
		f17: Number of times diabetic peripheral circulation disease codes were assigned

	Diagnosis note	f18: Number of times clinician's notes containing type 2 diabetes
		f19: Number of times clinician's notes containing diabetes but the type was not specified
		f20: Number of times clinician's notes containing diabetic retinopathy
		f21: Number of times clinician's notes containing diabetic neuropathy
		f22: Number of times clinician's notes containing diabetic eye disease
		f23: Number of times clinician's notes containing diabetic kidney disease
		f24: Number of times clinician's notes containing diabetic cerebral vascular disease
		f25: Number of times clinician's notes containing diabetic peripheral circulation disease
Outpatient diagnosis record	Diagnosis code	f26: Number of times type 2 diabetes codes were assigned
		f27: Number of times diabetes codes were assigned, but the type of diabetes is not specified
		f28: Number of times diabetic retinopathy codes were assigned
		f29: Number of times diabetic neuropathy codes were assigned
		f30: Number of times diabetic eye disease codes were assigned
		f31: Number of times diabetic kidney disease codes were assigned
		f32: Number of times diabetic cerebral vascular disease codes were assigned
		f33: Number of times diabetic peripheral circulation disease codes were assigned
	Diagnosis note	f34: Number of times clinician's notes containing type 2 diabetes
		f35: Number of times clinician's notes containing diabetes but the type was not specified
		f36: Number of times clinician's notes containing diabetic retinopathy
		f37: Number of times clinician's notes containing diabetic neuropathy
		f38: Number of times clinician's notes containing diabetic eye disease
		f39: Number of times clinician's notes containing diabetic kidney disease
		f40: Number of times clinician's notes containing diabetic cerebral vascular disease
		f41: Number of times clinician's notes containing diabetic peripheral circulation disease
Inpatient discharge summary	Diagnosis note	f42: Number of times summary notes containing type 2 diabetes
		f43: Number of times summary notes containing diabetes but the type was not specified
		f44: Number of times summary notes containing diabetic retinopathy

		f45: Number of times summary notes containing diabetic neuropathy
		f46: Number of times summary notes containing diabetic eye disease
		f47: Number of times summary notes containing diabetic kidney disease
		f48: Number of times summary notes containing diabetic cerebral vascular disease
		f49: Number of times summary notes containing diabetic peripheral circulation disease
Inpatient diagnosis record	Diagnosis code	f50: Number of times type 2 diabetes codes were assigned
		f51: Number of times diabetes codes were assigned, but the type of diabetes is not specified
		f52: Number of times diabetic retinopathy codes were assigned
		f53: Number of times diabetic neuropathy codes were assigned
		f54: Number of times diabetic eye disease codes were assigned
		f55: Number of times diabetic kidney disease codes were assigned
		f56: Number of times diabetic cerebral vascular disease codes were assigned
		f57: Number of times diabetic peripheral circulation disease codes were assigned
	Diagnosis note	f58: Number of times clinician's notes containing type 2 diabetes
		f59: Number of times clinician's notes containing diabetes but the type was not specified
		f60: Number of times clinician's notes containing diabetic retinopathy
		f61: Number of times clinician's notes containing diabetic neuropathy
		f62: Number of times clinician's notes containing diabetic eye disease
		f63: Number of times clinician's notes containing diabetic kidney disease
		f64: Number of times clinician's notes containing diabetic cerebral vascular disease
		f65: Number of times summary notes containing diabetic peripheral circulation disease
Prescription record	Medication	f66: Number of prescriptions for oral hypoglycemic
		f67: Number of prescriptions for insulin
		f68: Number of prescriptions for Chinese traditional hypoglycemic
		f69: Number of prescriptions for a mixture of western and Chinese traditional oral hypoglycemic
		f70: Number of prescriptions for Epalrestat

		f71: Number of prescriptions for Alpha-glucosidase inhibitor
		f72: Number of prescriptions for Dipeptidylpeptidase IV(DPP-IV) inhibitors
		f73: Number of prescriptions for Meglitinides
		f74: Number of prescriptions for Sulfonylureas
		f75: Number of prescriptions for Thiazolidinediones
		f76: Number of prescriptions for Biguanides
		f77: Number of prescriptions for Incretin Mimetics
		f78: Number of prescriptions for GLP-1 (glucagon-like peptide 1) mimetics
		f79: Number of prescriptions for compounds of sulfonylurea and thiazolidinedione
		f80: Number of prescriptions for compounds of Biguanides and Dipeptidylpeptidase IV(DPP-IV) inhibitors
		f81: Number of prescriptions for compounds of Biguanides and Sulfonylureas compounds
		f82: Number of prescriptions for compounds of Biguanides and Thiazolidinediones
Laboratory test reports	Venous plasma glucose test	f83: Number of times for 2-hours plasma glucose tests
		f84: Number of times for 2-hours plasma glucose tests≥11.1mmol/l (200mg/dl)
		f85: The maximum value of 2-hours plasma glucose tests
		f86: The minimum value of 2-hours plasma glucose tests
		f87: The number of times for fasting plasma glucose tests
		f88: The number of times for fasting plasma glucose tests ranging from 6.1 to 7.0 mmol/l (110 and 126 mg/dl)
		f89: The maximum value of fasting plasma glucose tests
		f90: The minimum value of fasting plasma glucose tests
		f91: Number of times for random plasma glucose tests
		f92: Number of times for random plasma glucose tests≥11.1mmol/l (200mg/dl)
		f93: The maximum value of random plasma glucose tests
		f94: The minimum value of random plasma glucose tests
	Peripheral plasma	f95: Number of times for 2-hours peripheral plasma glucose tests
		f96: Number of times for 2-hours peripheral plasma glucose tests ≥ 11.1mmol/l (200mg/dl)

	glucose test	f97: The maximum value of 2-hours peripheral plasma glucose tests
		f98: The minimum value of 2-hours peripheral plasma glucose tests
		f99: Number of times for peripheral fasting plasma glucose tests
		f100: Number of times for peripheral fasting plasma glucose tests ranging from 6.1 to 7.0 mmol/l (110 and 126 mg/dl)
		f101: The maximum value of peripheral fasting plasma glucose tests
		f102: The minimum value of peripheral fasting plasma glucose tests
		f103: Number of times for random peripheral plasma glucose tests
		f104: Number of times for random peripheral plasma glucose tests ≥ 11.1mmol/l (200mg/dl)
		f105: The maximum value of random peripheral plasma glucose tests
		f106: The minimum value of random peripheral plasma glucose tests
	HbA1C test	f107: Number of times for HbA1c tests
		f108: Number of times for HbA1C tests ≥ 6.5%
		f109: The maximum value of HbA1C tests
		f110: The minimum value of HbA1C tests

Appendix B: A list of diabetic medicine

Medicine is a principal factor to characterize phenotypes of subjects with type 2 diabetes mellitus (T2DM). In this paper, we use prescribed medicine listed in Table A2 as one of our seven sources to construct medicine related features as listed in Table A1.

Table A2. A list of medicine associated with subjects with type 2 diabetes mellitus

Category of medicine	Chinese generic name	Translated English generic name
Western Medicine	依帕司他	Epalrestat (A medicine treating for diabetic neuropathy)
	阿卡波糖	Acarbose (Alpha-glucosidase inhibitor)
	伏格列波糖	Voglibose (Alpha-glucosidase inhibitor)
	米格列醇	Miglitol (Alpha-glucosidase inhibitor)
	利拉利汀	Linagliptin (Dipeptidylpeptidase IV(DPP-IV) inhibitors)
	沙格列汀	Saxagliptin (Dipeptidylpeptidase IV(DPP-IV) inhibitors)
	维格列汀	Vidagliptin (Dipeptidylpeptidase IV(DPP-IV) inhibitors)
	西格列汀	Sitagliptin (Dipeptidylpeptidase IV(DPP-IV) inhibitors)
	那格列奈	Nateglinide (Meglitinides)
	瑞格列奈	Regalinide (Meglitinides)
	醋酸己脲	Acetohexamide (Sulfonylureas)
	格列本脲	Glyburide (Sulfonylureas)
	格列吡嗪	Glipizide (Sulfonylureas)
	格列喹酮	Gliquidone (Sulfonylureas)
	格列美脲	Glimepiride (Sulfonylureas)
	格列齐特	Gliclazide (Sulfonylureas)
	甲苯磺丁脲	Tolbutamide (Sulfonylureas)
	氯磺丙脲	Chlorpropamide (Sulfonylureas)
	马来酸罗格列酮和格列美脲	Glimepiride and rosiglitazone
	西格列汀二甲双胍片	Metformin and sitagliptin
	二甲双胍格列吡嗪	Metformin and glipizide
	格列本脲盐酸二甲双胍	Metformin and glyburide
	吡格列酮二甲双胍	Metformin and pioglitazone
	二甲双胍马来酸罗格列酮片	Metformin and rosiglitazone

	吡格列酮	Pioglitazone (Thiazolidinediones)
	罗格列酮	Rosiglitazone (Thiazolidinediones)
	曲格列酮	Troglitazone (Thiazolidinediones)
	苯乙双胍	Phenformin (Biguanides)
	二甲双胍	Metformin (Biguanides)
	普兰林肽	Pramlintide (Incretin Mimetics)
	艾塞那肽	Exenatide symthetic (GLP-1(glucagon-like peptide 1) mimetics)
	利拉鲁肽	Liraglutide (GLP-1(glucagon-like peptide 1) mimetics)
	利西拉来	Lixisenatide (GLP-1 (glucagon-like peptide 1) mimetics)
Integration of Traditional Chinese Medicine and Western Medicine	葛根消渴丸	XiaoKeWan (The Root of Kudzu Vine)
	地黄消渴丸	XiaoKeWan (Radices Rehmanniae)
	黄芪消渴丸	XiaoKeWan (Astragalus Mongholicus)
	天花粉消渴丸	XiaoKeWan (Radix Trichosanthis)
	玉米须消渴丸	XiaoKeWan (Stigmata Maydis)
	南五味子消渴丸	XiaoKeWan (Kadsura Longepedunculata)
	山药消渴丸	XiaoKeWan (Chinese Yam)
	格列本脲消渴丸	XiaoKeWan (Glibenclamide)
Traditional Chinese Medicine	参花消渴茶	ShenHuaXiaoKeCha (Ginseng, Astragalus Mongholicus, The Root of Kudzu Vine, Rhizoma Anemarrhenae, Radix Trichosanthis, Cortex Lycii Radicis, Radix Polygonati Officinalis, Green Tea, Rhizoma Phragmitis, Carthamus Tinctorious, The Dodder Weed, Gypsum, Platycodon Grandiflorum)
	参芪降糖	ShenQiJiangTang (Panax Ginseng Leaves Extract,The Fruit of Chinese Magnoliavine, Astragalus Mongholicus, Chinese Yam, Radices Rehmanniae, Fructus Rubi, Radix Ophiopogonis, Poria Cocos, Radix Trichosanthis, The Rhizome of Oriental Water Plantain, The Fruit of Chinese Wolfberry)
	地骨降糖	DiGuJiangTang (Radix Curcumae, Cortex Lycii Radicis, Fructus Perillae, Tortoise Shell, Lumbricus, Leech, Cordyceps Sinensis)
	甘露消渴	GanLuXiaoKe (Prepared Rehmannia Root, Radices Rehmanniae, Cortex Lycii Radicis, Ginseng, The Fruit of Chinese Wolfberry, Astragalus Mongholicus, The Dodder Weed, Fructus Corni, Codonopsis Pilosula, Coptis Chinensis)
	降糖甲	JiangTangJia (Astragalus Mongholicus, Rhizoma Polygonati, Radices Rehmanniae, Radix Pseudostellariae, Radix Trichosanthis, Ginseng, Chinese yam, Gypsum, Rhizoma Anemarrhenae, Astragalus Mongholicus, Radix Trichosanthis, Poria Cocos, Radix Ophiopogonis, Radix Rehmanniae Recens, Cortex Lycii Radicis, Stigmata Maydis, Fructus Corni, Liquorice)
	降糖宁	JiangTangNing (Ginseng, Chinese yam, Gypsum, Rhizoma Anemarrhenae, Astragalus Mongholicus, Radix Trichosanthis, Poria Cocos, Radix Ophiopogonis, Radix Rehmanniae Recens, Cortex Lycii Radicis, Stigmata Maydis, Fructus Corni, Liquorice)
	降糖舒胶囊	JiangTangShuJiaoNang (Ginseng, The Fruit of Chinese Wolfberry, Astragalus Mongholicus, Radix et Caulis Acanthopanacis Senticosi, Rhizoma Polygonati, Semen Amomi Amari, Concha Ostreae, Radices Rehmanniae, Prepared Rehmannia Root, The Root of Kudzu Vine, The Root of Red-Rooted Salvia, Semen Litchi, Rhizoma Anemarrhenae, Gypsum,

		Semen Euryales, Chinese Yam, Radix Scrophulariae, The Fruit of Chinese Magnoliavine, Radix Ophiopogonis, The Root of Three-nerved Spicebush, Radix Trichosanthis, Fructus Aurantii)
	金芪降糖	JinQiJiangTang (Pearl, Astragalus Mongholicus, Rhizoma Polygonati, Scutellaria Baicalensis, Radices Rehmanniae, Radix Trichosanthis, Radix Ophiopogonis, Dendrobe, Cicada Slough, Endothelium Corneum Gigeriae Galli, Chinese Yam, Semen Astragali Complanati, Pericarpium Citri Reticulatae Viride, The Root of Kudzu Vine)
	晶珠糖尿康	JingZhuTangNiaoKang (Fructus Chebulae, Carthamus Tinctorious, Amomum Kravanh, Rock Extract, Shellac, Radix Et Rhizoma Rubiae, Fructus Phyllanthi, Turmeric, Berberis Kansuensis Schneid, Tribulus Terrestris L., Lapis Micae Aureus, Juniperus Formosana, Saxifraga Umbellulata Hook. f. et Thoms, Corydalis Impatiens, Leguminosae, Bear Gall, Bos Taurus Domesticus Gmelin)
	渴乐宁	KeLeNing (Astragalus Mongholicus, Rhizoma Polygonati, Radices Rehmanniae, Radix Pseudostellariae, Radix Trichosanthis)
	糖脉康	TangMaiKang (Astragalus Mongholicus, Radix Rehmanniae Recens, The Root of Red-rooted Salvia, The Root of Kudzu Vine, Folium Mori, Herba Epimedii)
	糖尿乐	TangNiaoLe (Radix Trichosanthis, Radix Ginseng Rubra, Chinese Yam, Astragalus Mongholicus, Radices Rehmanniae, The Fruit of Chinese Wolfberry, Rhizoma Anemarrhenae, Fructus Corni, The Root of Kudzu Vine, The Fruit of Chinese Magnoliavine, Radix Asparagi, Poria Cocos, Endothelium Corneum Gigeriae Galli)
	糖脂消	TangZhiXiao (Astragalus Mongholicus, The Root of Red-rooted Salvia, Stephania Tetrandra, Cortex Lycii Radicis, Coptis Chinensis, Bighead Atractylodes Rhizome)
	洗胰清糖素	XiYiQingTangSu (Folium Mori, The Root of Kudzu Vine, Balsam Pear, Radix Polygonati Officinalis)
	消渴康	XiaoKeKang (Gypsum, Rhizoma Anemarrhenae, Radix Rehmanniae Recens, Radix Ophiopogonis, Radix Trichosanthis, Radix Polygonati Officinalis, Radix Scrophulariae, The Root of Bidentate Achyranthes, The Root of Red-rooted Salvia, The Rhizome of Oriental Water Plantain, Codonopsis Pilosula, Fructus Corni, Folium Eriobotryae, Kadsura Longepedunculata)
	消渴灵片	XiaoKeLing Pian (Radices Rehmanniae, The Fruit of Chinese Magnoliavine, Radix Ophiopogonis, Cortex Moutan Radicis, Astragalus Mongholicus, Coptis Chinensis, Poria Cocos, Radix Ginseng Rubra, Radix Trichosanthis, Gypsum, The Fruit of Chinese Wolfberry)
	玉泉丸	YuQuanWan (The Root of Kudzu Vine,Radix Trichosanthis, Radices Rehmanniae, Radix Ophiopogonis, The Fruit of Chinese Magnoliavine, Liquorice)
	珍芪降糖	ZhenQiJiangTang (Pearl, Astragalus Mongholicus, Rhizoma Polygonati, Scutellaria Baicalensis, Radix Rehmanniae Recens, Radix Trichosanthis, Radix Ophiopogonis, Dendrobe, Cicada Slough, Endothelium Corneum Gigeriae Galli, Chinese Yam, Semen Astragali Complanati, Pericarpium Citri Reticulatae Viride, The Root of Kudzu Vine)

Appendix C: A list of positive and negative diagnosis notes related with T2DM

Diagnosis notes existing in diagnosis reports or clinical summaries are represented as unstructured texts. We create a dictionary of diagnosis notes related with T2DM. There are two types of diagnosis notes: positive and negative. We assume that if a subject's EHR data contains positive diagnosis notes, but not negative diagnosis notes, then the positive diagnosis notes are counted to construct features associated with diagnosis notes.

Table A3. A list of positive and negative diagnosis notes related with T2DM

Diagnosis note category	**Chinese notes**	**Translated English notes**
Positive diagnosis notes	2 型糖尿病	Type 2 diabetes
	2-糖尿病	
	2 型糖尿病	
	2-型糖尿病	
	2 型糖尿病	
	Ⅱ型糖尿病	
	II 型糖尿病	
	II 糖尿病	
	II 型糖尿病	
	二型糖尿病	
	糖尿病 II 型	
	糖尿病（Ⅱ型）	
	糖尿病 2	
	糖尿病 2 型	

	糖尿病Ⅱ型	
	糖尿病 II	
	糖尿病 II 型	
	非胰岛素依赖型糖尿病	Noninsulin-dependent diabetes mellitus
	糖尿病	Diabetes mellitus
Negative diagnosis notes	排除糖尿病	Exclusion of diabetes
	非糖尿病	
	糖尿病的特殊筛查	Special screening for diabetes
	糖尿病特殊筛查	
	糖尿病母亲的婴儿综合征	Syndrome of infant of diabetic mother
	糖尿病母亲的婴儿综合征	
	母亲伴妊娠糖尿病的婴儿综合征	
	妊娠糖尿病母亲婴儿综合征	
	糖尿病家族史	Family history of diabetes mellitus
	潜伏性糖尿病	Occult diabetes
	早期型糖尿病	Early type diabetes
	隐性糖尿病	Latent diabetes
	化学性糖尿病	Chemical diabetes
	糖尿病前期	Prediabetes
	胰岛素和口服降血糖[抗糖尿病]药中毒	Oral hypoglycemic drug poisoning
	口服降血糖[抗糖尿病]药中毒	

Appendix D: A list of 36 features summarized from 110 features as listed in Table A1

Features listed in Table A1 are extracted from seven sources, however, several features across sources are correlated. For instance, diagnosis-code related features appearing in "*communication report*", "*outpatient diagnosis record*" and "*inpatient diagnosis record*" are similar. These features have the same definition in above three sources, so they can be summarized as a new feature. In this way, eight new features (f'_{10} to f'_{17}) in the category of diagnosis codes as shown in Table A4 are summarized from 24 features (f_{10} to f_{17}, f_{26} to f_{33}, f_{50} to f_{57}) from Table A1. By using the same way, we summarize 32 similar diagnosis-note related features appearing in "*communication report*" (f_{18} to f_{25}), "*outpatient diagnosis record*" (f_{34} to f_{41}), "*inpatient diagnosis record*" (f_{42} to f_{49}) and "*inpatient discharge summary*" (f_{58} to f_{65}) into 8 new features (f'_{18} to f'_{25}) in the category of diagnosis notes as shown in Table A4.

Features as listed in the "*laboratory test report*" of Table A1 are also correlated with each other. For instance, features ranging from f_{83} to f_{86} are all correlated with venous 2-hours plasma glucose test. In order to reduce negative influences of correlated features on the performances of classification models such as k nearest neighbors, we only keep features which are positive signals of type 2 diabetes. For instance, feature f_{84} characterizing the number of times 2-hours plasma glucose test≥11.1mmol/l, which is a positive signal of type 2 diabetes conditions. So do feature f_{88}, f_{92}, f_{96}, f_{100}, f_{104} and f_{108}.

Most of subjects only take a small number of medicine listed in Table A2, as a result, the data covering features ranging from f_{66} to f_{82} has a big sparsity, which will influence the performances of computational models to learn patterns of T2DM [1]. In order to avoid a big sparsity, we transform original features ranging from f_{66} to f_{69} into new ones ranging from f'_{26} to f'_{29} as shown in Table A4.

Table A4. The original 110 constructed features as shown in Table A1 are transformed into 36 features via summarizing similar features across seven sources: "*communication report*", "*outpatient diagnosis record*", "*inpatient diagnosis record*", "*inpatient discharge summary*", "*prescription report*" and "*laboratory report*".

Category of features	New Merged Feature
Demographic information	f'1 = f1 f'2 = f2 f'3 = f3
Self-reporting notes	f'4 = f4 f'5 = f5 f'6 = f6 f'7 = f7 f'8 = f8 f'9 = f9
Diagnose codes	f'10=f10+f26+f50 f'11=f11+f27+f51 f'12=f12+f28+f52 f'13=f13+f29+f53 f'14=f14+f30+f54 f'15=f15+f31+f55

	f’16=f16+f32+f56 f’17=f17+f33+f57
Diagnose notes	f’18=f18+f34+f42+f58 f’19=f19+f35+f43+f59 f’20=f20+f36+f44+f60 f’21=f21+f37+f45+f61 f’22=f22+f38+f46+f62 f’23=f23+f39+f47+f63 f’24=f24+f40+f48+f64 f’25=f25+f41+f49+f65
Medication	f’26 = f66 f’27 = f67 f’28 = f68 f’29 = f69
Plasma glucose and HbA1C tests	f’30=f84 f’31=f88 f’32=f92 f’33=f96 f’34=f100

	f’35=f104 f’36=f108

Appendix E: A list of 8 features summarized from 36 features as listed in Table A4

36 features in Table A4 are summarized as 8 features in following 6 categories:

(1) **Patients' demographic information**: ranging from f''_1 to f''_3.

(2) **Self-report**: summarize 6 features ranging from f'_4 to f'_9 in Table A4 as f''_4 in Table A5 to represent the total number of times diabetic phenomena such as body weight loss, persistent hunger, polyuria, polydipsia, prescribed diabetes medicine and returning visits for diabetes were reported by subjects in the source of "*communication report*".

(3) **Diagnosis code**: summarize 8 features ranging from f'_{10} to f'_{17} in Table A4 as f''_5 in Table A5 to represent the total number of times diabetic diagnosis-codes are assigned to a subject in "*communication report*", "*outpatient diagnosis record*" and "*inpatient diagnosis report*".

(4) **Diagnosis note**: summarize 8 features ranging from f'_{18} to f'_{25} in Table A4 as f''_6 in Table A5 to represent the total number of times diabetic diagnosis-notes are described in a subject's "*communication report*", "*outpatient diagnosis record*", "*inpatient diagnosis record*" and "*inpatient discharge summary*".

(5) **Medication**: summarize 4 features ranging from f'_{26} to f'_{29} in Table A4 as f''_7 in Table A5 to represent the total number of times diabetic medicines as listed in Table A2 are prescribed in a subject's prescription record.

(6) **Plasma glucose and HbA1C test**: summarize 7 features ranging from f'_{30} to f'_{36} in Table A4 as f''_8 in Table A5 to represent the total number of times venous plasma glucose, peripheral plasma glucose (fasting plasma glucose ≥126 mg/dl or 2-hours plasma glucose≥200 mg/dl or random plasma glucose ≥200 mg/dl) and HbA1C tests are abnormal.

Table A5. The 8 features after summarizing related features within a category such as "self-reporting note", "diagnosis code", "diagnosis note", "medication", "plasma glucose" and "HbA1C test".

Category of features	Feature
Demographic information	f''1 = f'1 f''2 = f'2 f''3 = f'3;
Self-reporting note	f''4 = f'4+ f'5+ f'6+ f'7+ f'8+ f'9
Diagnosis code	f''5 = f'10+ f'11+ f'12+ f'13+ f'14+ f'15+ f'16+ f'17
Diagnosis note	f''6 = f'18+ f'19+ f'20+ f'21+ f'22+ f'23+ f'24+ f'25
Medication	f''7 = f'26+ f'27+ f'28+ f'29
Plasma glucose and HbA1C test	f''8 = f'30+ f'31+ f'32+ f'33+ f'34+ f'35+ f'36

Appendix F: Expert algorithm for the identification of subjects with T2DM

The expert algorithm[2] we used as our baseline to do performance comparisons is depicted in Figure A1. The performance of the algorithm had been successfully validated at multiple eMERGE Network[3] sites in the USA. The algorithms utilized various types of information including diagnosis codes, medication orders, laboratory results and clinical notes. We applied this algorithm on all of our investigated EHR sources including diagnoses, laboratory results, medications, communication reports and clinical notes. Notably, the expert algorithm and our approach both used the same EHR sources.

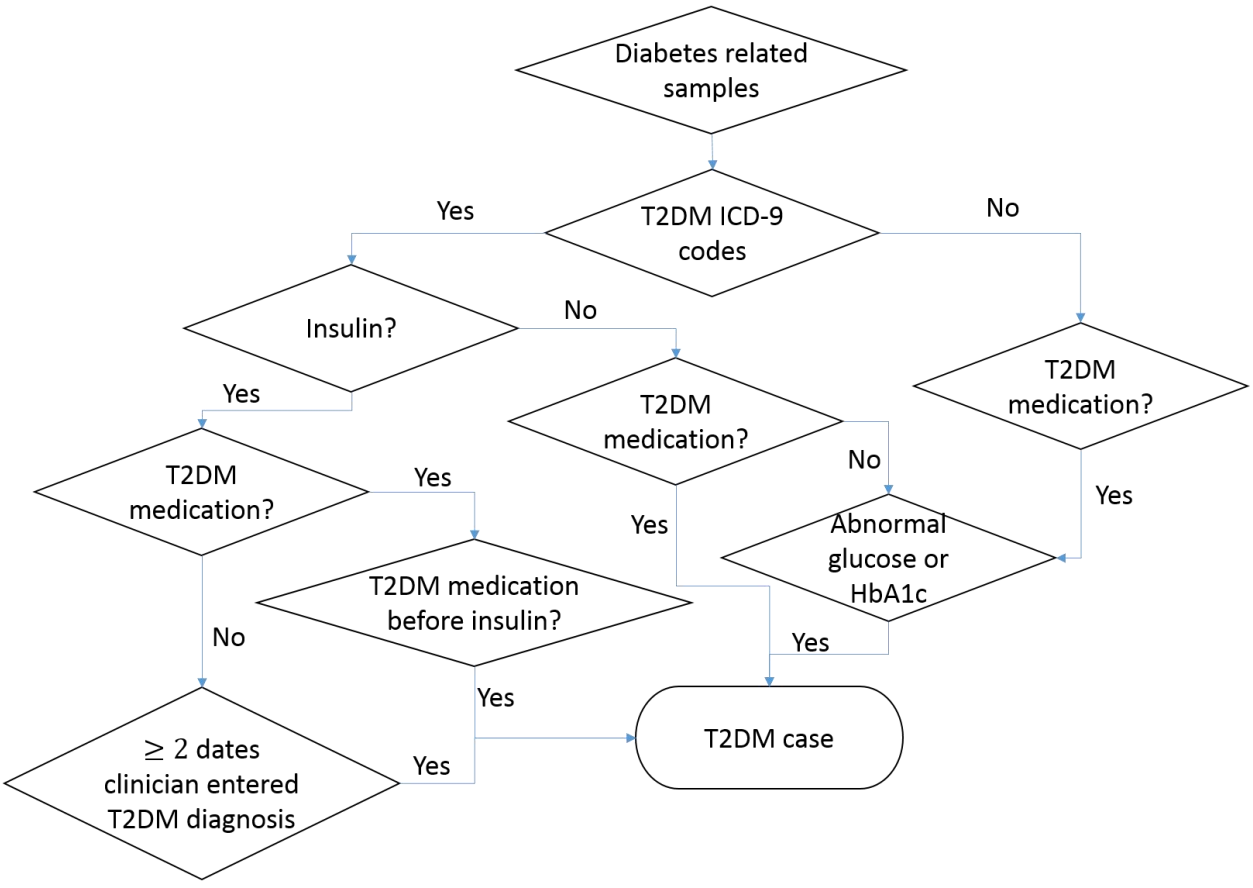

Figure A1. Expert algorithm for the identification of subjects with T2DM

Reference

1. B. K. Natarajan**.** Sparse Approximate Solutions to Linear Systems. SIAM J. Comput. 1995; 24(2):227–234.
2. A. N. Kho, et.al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association. 2012; 19(2):212-218
3. McCarty CA, et.al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Medical Genomics. 2011; 4:13